University of Southampton

University of Southampton Research Repository

Copyright © and Moral Rights for this thesis and, where applicable, any accompanying data are retained by the author and/or other copyright owners. A copy can be downloaded for personal non-commercial research or study, without prior permission or charge. This thesis and the accompanying data cannot be reproduced or quoted extensively from without first obtaining permission in writing from the copyright holder/s. The content of the thesis and accompanying research data (where applicable) must not be changed in any way or sold commercially in any format or medium without the formal permission of the copyright holder/s.

When referring to this thesis and any accompanying data, full bibliographic details must be given, e.g.

Thesis: Author (Year of Submission) "Full thesis title", University of Southampton, name of the University Faculty or School or Department, PhD Thesis, pagination.

Data: Author (Year) Title. URI [dataset]

University of Southampton

Faculty of Medicine

Human Development and Health

Integration of health informatics: 'big data' for clinical translation in inflammatory bowel disease

by

Imogen Siân Stafford

ORCID ID 0000-0003-1666-1906

Thesis for the degree of Doctor of Philosophy

June 2023

University of Southampton

Abstract

Faculty of Medicine

Human Development and Health

Thesis for the degree of Doctor of Philosophy

Integration of health informatics: ‘big data’ for clinical translation in inflammatory bowel disease

by

Imogen Siân Stafford

Inflammatory bowel disease (IBD) is a chronic, complex autoimmune disease characterised by relapsing-remitting gastrointestinal tract inflammation. It is considered to arise from interactions between an individual’s genetic susceptibility, environmental factors, immune dysregulation, and gut microbial dysbiosis. Genetics can make a larger contribution to IBD pathology in some patients, and this is thought to be linked to age of diagnosis, with genetic factors having the largest effects in very young children. There are two main subtypes of IBD: ulcerative colitis (UC) and Crohn’s disease (CD). Within subtypes, there are different disease behaviours and severities. One particular disease behaviour of interest is the stricturing endotype, which causes a narrowing of the gastrointestinal tract that often requires surgery.

This thesis first examines oxidative stress in IBD patients, through the use of assay data. Here, statistical and machine learning (ML) methods are employed to examine the relationship between clinical and genomic characteristics of a set of paediatric patients, and their measured oxidative stress and antioxidant potential. In this work, no results suggested that these assay data could be used as an indicator for these clinical features, or for pathogenic variation in key oxidative stress genes.

The predominant focus of this thesis is the use of genomic data and ML to stratify IBD patients. In order to prepare genomic data for use in ML pipelines, the GenePy algorithm was used. GenePy takes in information regarding zygosity, allele frequency, and predicted deleteriousness for every variant in a gene. The scores for each variant are summed to create an overall gene score, and this becomes are per-gene, per-individual matrix of scores. The two clinical problems analysed here were classifying IBD patients according to their subtype, and stratifying CD patients by the presence or absence of a stricturing endotype. This was achieved with an ML random forest classifier. Optimisation of both the input data and ML algorithm for these classifications was a important aspect of this work. Several gene panels were trialled for these classifications, and an autoimmune gene panel outperformed an IBD gene panel for determining IBD subtype. Stratifying CD patients by their stricturing endotype was subsequently performed with a random survival

forest, which combined a random forest with survival analysis methods. This method is better suited to the longitudinal nature of stricturing endotype developed. This work demonstrated challenges that arise from the sparsity of genomic data, and required the development of a pipeline that could reduce the sparsity of the features used by the ML algorithm.

The patient stratification performed here demonstrated strong evidence for the presence of different genomic variation patterns within IBD subtypes, and within the CD stricturing endotype. With increased dataset sizes, it may be possible to more clearly detect and cluster patients according to their genomic variation. In order to take full advantage of this knowledge, there is an additional requirement for deep, varied and longitudinal clinical data. Then, genomic data can guide each patient's clinical pathway, providing individuals with more personalised, life-long care.

Table of Contents

Table of Tables

Table of Figures

List of Accompanying Materials

All accompanying materials can be found at the following doi:
https://doi.org/10.5258/SOTON/D2655

Chapter 3:

- Python and shell scripts required to execute the joint calling pipeline and GenePy 1.3 pipeline detailed in Sections 3.3.4 and 3.3.5.
- Accompanying static GitHub pages providing further instruction and detail for the joint calling pipeline and GenePy 1.3 pipeline detailed in Sections 3.3.4 and 3.3.5.
- Python and shell scripts required to execute the updated joint calling pipeline and GenePy 1.4 pipeline detailed in Section 3.4.
- Accompanying static GitHub pages providing further instruction and detail for the joint calling pipeline and GenePy pipeline detailed in Section 3.4.

Chapter 4:

- R scripts for the supervised machine learning of genomic data and oxidative stress and antioxidant potential assay data

Chapter 5 and Chapter 6:

- Quality control report for the IBD cohort
- Remapped list of highly mutable genes, from Fuentes Fajardo et al.
- Gene panels utilised in machine learning: autoimmune disease gene panel, inflammatory bowel disease monogenic genes and GWAS genes panel, stricturing endotype gene panels, and NOD-signalling pathway gene panel. Some of these panels are also used in Chapter 7.
- Python scripts for machine learning for classification of inflammatory bowel disease subtype, the Crohn’s disease stricturing endotype, and age of onset classifiers
- Genes identified in a literature review that are associated with the Crohn’s disease stricturing endotype, and their source.

Chapter 7:

- Python script of random survival forest for stratification of Crohn’s disease patients by stricturing endotype

- Results for the top features selected during Bayes Search and Grid Search trials to monitor feature selection stability.

Research Thesis: Declaration of Authorship

Print name: IMOGEN SIAN STAFFORD

Title of thesis: **Integration of health informatics: 'big data' for clinical translation in inflammatory bowel disease**

I declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:-

   Stafford IS, Mossotto E, Ashton JJ, Cheng G, Beattie RM, Ennis S. Supervised machine learning classifies inflammatory bowel disease patients by subtype using whole exome sequencing data. J Crohns Colitis. May 2023. https://doi.org/10.1093/ecco-jcc/jjad084

   Stafford IS, Gosink MM, Mossotto E, Ennis S, Hauben M. A systematic review of artificial intelligence and machine learning applications to inflammatory bowel disease, with practical guidelines for interpretation. *Inflamm. Bowel Dis*. June 2022.

   Stafford IS, Kellermann M, Mossotto E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ Digit Med*. Mar 2020. doi: 10.1038/s41746-020-0229-3.

Signature: .... Date: 08 June 2023

Acknowledgements

Firstly, I would like to acknowledge Prof. Sarah Ennis for her role as main supervisor. Her continued support, both professionally and personally, has been invaluable.

I'd also like to thank my other supervisors Dr Enrico Mossotto, Prof. Mark Beattie, and Prof. Benjamin MacArthur for sharing their knowledge, and their guidance.

I'm grateful to Prof. Martin Feelisch and Prof. Mahesan Niranjan for lending their expertise in their respective fields.

I'd like to acknowledge all the patients and their families that have taken part in the Genetics of Inflammatory Bowel Disease study. This research is not possible without them. I'd also like to thank the research nurses who recruited the patients to this study: Rachel Haggarty, Rachel Brampton, and Genevieve Roberts.

Thank you to Nikki Graham for sample processing and extraction of DNA, and Florina Borca and Hang Phan of the Southampton Biomedical Research Centre data science team who helped with clinical data extraction.

I'd like to acknowledge the funders: University of Southampton Institute for Life Sciences, and the NIHR Southampton Biomedical Research Centre.

Thank you to my colleagues on the Genetics of Inflammatory Bowel Disease study Dr James Ashton and Dr Guo Cheng, as well as colleagues past and present in the Human Genetics and Genomics group: Dr Ellie Seaby, Dr Gary Leggatt, Dr Clare Horscroft and Dr Carolina Jaramillo Oquendo. Everyone has provided moral support and many laughs during my candidature.

Finally, I am so grateful to my family for supporting me.

COVID-19 Impact Statement

The Genetics of Inflammatory Bowel Disease study, my primary source of data for this project, was paused in March 2020, and was reopened in August 2020. Thankfully, as the study has been running for over a decade, there was already a considerable amount of data from the study available for my research. We also continued to receive new data during 2020. As my research was focussed on bioinformatics and computational methods, progressing the project was minimally impacted by COVID-19.

In addition, COVID-19 did impact me personally. Restrictions put in place in healthcare settings meant a lengthy wait for tests and a new diagnosis and treatment of a long-term health condition, the symptoms of which did affect my ability to progress the research.

List of commonly used abbreviations

AI	Artificial intelligence
ANOVA	One-way analysis of variance
AUC	Area under the receiver-operator curve
BED	Browser extensible data
BWA	Burrows-Wheeler Aligner
CADD	Combined Annotation-Dependent Depletion
CAGI	Critical Assessment of Genome Interpretation
CD	Crohn’s disease
C-index	Concordance index
CPH	Cox Proportional Hazards
CRP	C-reactive protein
CV	Cross validation
DNA	Deoxyribonucleic acid
Ensembl-VEP	Ensembl Variant Effect Predictor
FRAP	Ferric reducing ability of plasma
GATK	Genome Analysis Toolkit
gnomAD	Genome Aggregation Database
GQ	Genotype quality
GVCF	Genotyped variant call format
GWAS	Genome-wide association studies
IBD	Inflammatory Bowel Disease
IBDU	Inflammatory Bowel Disease Unclassified

JAK	Janus kinase
LASSO	Least Absolute Shrinkage and Selection Operator
ML	Machine learning
MRI	Magnetic resonance imaging
NGS	Next generation sequencing
NOD2	Nucleotide-binding oligomerization domain-containing protein 2
PC	Principal component
PCA	Principal component analysis
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RNA	Ribonucleic acid
ROS	Reactive oxygen species
RSF	Random survival forest
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
TBARS	Thiobarbituric acid reactive substances
TFT	Total free thiol
UC	Ulcerative colitis
VCF	Variant call format
VQSR	Variant quality score recalibration
VEOIBD	Very early onset inflammatory bowel disease
WES	Whole exome sequencing

Chapter 1 Introduction

Traditionally, inflammatory bowel disease (IBD) has been thought of as a multifactorial autoimmune disease that develops through the complex interactions between a person's genetics and the environment. Prior to the 21st century, little was known about the genetics of IBD, the major discovery being the *NOD2* gene's causative role [1-3]. In the last 20 years great strides have been made in the technologies that can be used to identify genes related to disease. IBD has been among those diseases that have benefitted from new genetic technologies. Now, over 200 genes are known to have a role in IBD pathology [4]. The subsequent development of high throughput sequencing allowed for closer examination of specific changes in the genetic code, and their possible links to the development of IBD. Using this sequencing technology a number of patients, usually diagnosed in very early childhood, were shown to have disease that is only caused by a change in one gene [5, 6], demonstrating that IBD is not always multifactorial.

Despite the new knowledge obtained through these technologies, there have been few changes to the clinical management of IBD patients. In the majority of cases, treatment is still based on clinical information gathered during investigation. Aside from the recent introduction of genetic sequencing for infants with a potential IBD diagnosis, a genetic investigation is not standard. One of the difficulties in implementing genetic investigations is that, although there is knowledge of IBD genetics, it is not known how these genes relate to specific patient phenotypes. Another obstacle is that these new technologies generate substantial amounts of data that are difficult and time-consuming to analyse. For these reasons, new methods are necessary.

Here, a combination of linear and non-linear methods are explored to facilitate translation of the genetic basis for patient's IBD into the clinic. Both paediatric and adult IBD patients are present in the IBD cohort that is utilised in the analysis. The aim is to stratify patients according to their observed genetics in order that their treatment and management is personalised. This should lead to better long-term outcomes for patients with this life-long disease.

1.1 Inflammatory bowel disease

Inflammatory bowel disease (IBD) is a complex autoimmune disease with two subtypes: Crohn's disease (CD) and ulcerative colitis (UC). The aetiology of IBD remains poorly understood, but four factors are known to contribute: genetics, environment, immune dysregulation, and gut microbial dysbiosis. This condition is characterised by chronic relapsing-remitting gastrointestinal tract inflammation, with location and pattern (continuous or discontinuous) of the inflammation differing depending on the IBD subtype [7, 8]. Information gained from histopathological investigations are also considered when determining the subtype [7, 8]. The symptoms of IBD are diverse, particularly in paediatric (<18 years) cases, and can include: diarrhoea, fever, fatigue, vomiting, anaemia, abdominal pain and growth retardation [7]. Extraintestinal manifestations are common in both adult and paediatric cases, and affects 25-35% of patients [9]. Accordingly, the Porto criteria [7] for paediatric IBD diagnosis, and the British Society of Gastroenterology consensus guidelines for adults, [8] utilises endoscopic and histopathological findings for diagnosis of CD or UC. If a diagnosis cannot be confirmed, the patient is diagnosed as IBD unclassified (IBDU). Individuals will often be diagnosed with CD or UC subsequent to an IBDU diagnosis [7, 8]. A correct diagnosis is imperative for patients, in order that they receive the correct treatments and interventions. Regardless of age at diagnosis, IBD is a chronic disease which requires lifelong monitoring and management by the patient and the clinician.

1.1.1 Epidemiology

The incidence of IBD surged during the late twentieth and early twenty-first century, particularly in western countries [10, 11]. High prevalence's of IBD have been reported in Canada [12] and Scotland [11] (0.7% and 0.8%, respectively). Approximately 25% of IBD presents during childhood [13], and research performed using data from the Wessex region in England explored incidence specifically among the paediatric population. A 50% increase in cases was observed from 2002 to 2012 (6.39/100,000 to 9.37/100,000), driven predominantly by CD cases [13]. The incidence in this region had increased to 10.54 per 100,000 by 2017 [14], and over 12 per 100,000 by 2021 [15]. For the overall IBD population in the United Kingdom, incidence has been estimated at 28.6 per 100,000 [16]. Although IBD was previously thought to be only predominant in western countries, incidence is documented to be rising in Asian and Latin American Countries. In Hong Kong, the last 20 years has seen IBD incidence rise from 1 to 3.1 per 100,000, while the highest documented incidence in the Asian-Pacific region is in India (9.31/100,000) [17]. The incidence of CD in Asia has risen quicker than UC [17]. Worldwide, the prevalence of IBD will only increase, compounding the health burden in the future.

1.1.2 Ulcerative colitis

Ulcerative colitis has a continuous inflammation pattern of the mucosa [18]. Inflammation is non-transmural and usually begins at the rectum and extends to some, or all, segments of the colon [18]. Patients are classified according to the extent of colonic involvement, and can be diagnosed with proctitis, left-sided colitis which involves the sigmoid colon and may or may not involve the descending colon, or pancolitis (Figure 1A) [18]. Pancolitis has been observed at a far higher frequency in child-onset versus adult-onset (82.2% versus 47.6%, children and adults respectively), and the reverse is true for proctitis (1.4% versus 17.0%, children and adults respectively) [19]. The most common symptoms of UC are bloody diarrhoea, rectal bleeding, weight loss and abdominal pain [9]. In paediatric cases, UC can be particularly difficult to diagnose due to heterogeneity in the UC phenotype [7]. The Porto criteria for paediatric patients describes the most reliable feature for UC diagnosis as colonic mucosal inflammation involving the rectum, but with no small bowel involvement, and no granulomas on biopsy [7]. There are five atypical UC presentations in paediatric disease: upper intestinal tract involvement, rectal sparing, short disease duration, left-sided colitis with an area of cecal inflammation, and acute severe UC that features more characteristics of CD (for example deep ulcers and transmural inflammation) [7].

1.1.3 Crohn’s disease

Where inflammation in UC is restricted to the colon, in CD inflammation can occur along the entire gastrointestinal tract, from the mouth to the anus (Figure 1B) [20]. Inflammation is transmural and patchy, but is more common in the terminal ileum, colon and ileocolon (47%, 28% and 21% respectively) than the upper gastrointestinal tract (3%) [18]. The symptoms of CD can often be more general, which results in a longer time to diagnosis [21]. A diagnosis of CD is made after considering clinical, radiographic, endoscopic and pathological findings [8]. CRP is a blood marker that is an initial indicator of an inflammatory disease, and is also used to monitor disease status [7]. Endoscopy can be used to access location of disease, and during this procedure a number of biopsies are often obtained for histopathological investigation [7, 8]. The extent of disease can be seen through the way in which the mucosa has been affected: from small ulcers in mild disease to large deep ulcers in a wavy pattern in severe disease [22]. Common biopsy findings in CD are discontinuous chronic inflammation, focal crypt distortion and granulomas [23]. CD patients can present with, or develop, complications such as strictures, fistulas and abscesses [20]. There can be confusion with UC in cases where inflammation is solely colonic [7]. In addition, some paediatric CD cases have been noted to present with isolated oral inflammation, and develop gastrointestinal luminal disease during their disease course [7]. It has been observed that paediatric CD is more common in male patients, and in adult patients this diagnosis is more common in females [19]. Additionally, isolated ileal and isolated colonic CD has been noted as more common in adult-onset disease, and paediatric onset CD was more likely to be extensive [19].

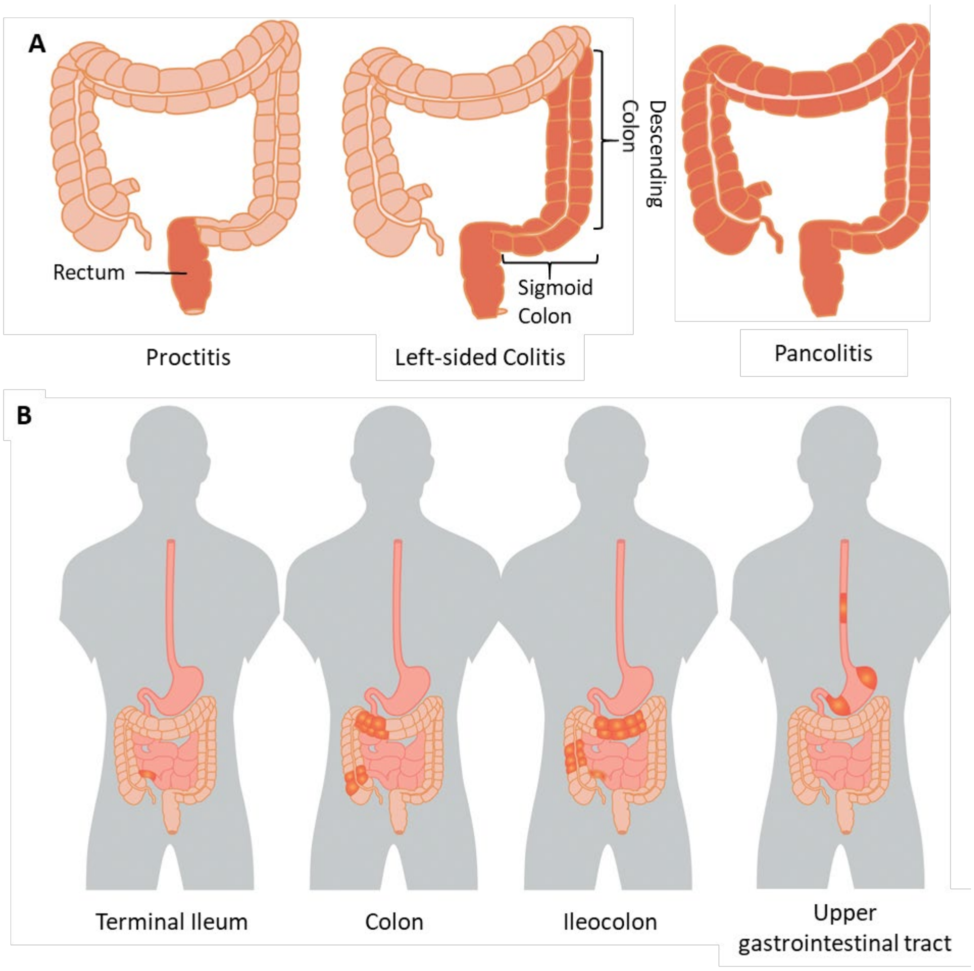


Figure 1 Locations of inflammation in UC and CD. A) The extent of inflammation in ulcerative colitis inflammation varies, and patients are classified accordingly. B) Different inflammation patterns in Crohn's disease. Image adapted from [24].

1.1.4 Inflammatory bowel disease unclassified

Where an individual exhibits phenotypic features of both CD and UC, they are given a diagnosis of inflammatory bowel disease unclassified (IBDU) [7, 8]. As previously mentioned there are cases where it can be difficult to distinguish between the main subtypes, and a common situation for a diagnosis of IBDU is inflammation is exclusive to the colon coinciding with CD presentations such as height delay, or other macroscopic and microscopic features [7]. An IBDU diagnosis can also be a result of an incomplete clinical investigation [25]. Additionally, paediatric-onset IBD is associated with an IBDU diagnosis, with a stronger likelihood of IBDU in infantile and very early onset (<6 years) patients [26]. A paediatric IBDU diagnosis can resolve to a diagnosis of CD or UC during disease course, but the rate of this diagnostic change is unclear. One study reports a diagnosis change in 32% of paediatric cohort, with a median follow up time of 5.7 years [25]. Another describes a change of diagnosis to UC or CD in 55% (median follow up 6.7 years) of paediatric cases [27].

1.1.5 Clinical classification systems

The complex presentation of the IBD subtypes can render decisions regarding appropriate therapies difficult. Treatment regimens are not only based on subtypes, but on disease behaviours, and their extent and severity. For these reasons guidelines for classifying patients were deemed important, and necessary. A system for classification of IBD sub-phenotypes for CD was first introduced in 1991, by The International Working Party in Rome [28]. This Rome classification system was based on anatomical distribution, operative history and clinical disease behaviour. The classification criteria evolved into the Vienna system in 1998, which was then based on age of onset, disease location and disease behaviour [28]. When this was revised into the Montreal classification the three criteria as per the Vienna classification remained, but adjustments were made within them [28]. The Montreal classification age of onset categories were a limitation in paediatric gastroenterology, as they were A1 under 16 years, A2 17-40 years and A3 above 40 years. A modification of the Montreal classification for paediatric patients called the Paris classification was introduced, which introduced age of onset categories A1a 0 to < 10 years, and A1b 10 to < 17 years [29]. The Paris classification also introduced amendments to categories within disease behaviour and location, and added a Growth category to document evidence of growth delay [29].

Sub-classification of UC was addressed in the Montreal classification and focused on the extent of the inflammation, and the severity in UC. These categories were modified in the Paris classification, with an additional class in extent of inflammation, and the use of the Paediatric Ulcerative Colitis Activity Index (PUCAI) for measuring severity [29]. The PUCAI consists of 6 categories where points are given based on the patient’s current status: abdominal pain, rectal bleeding, stool consistency and number, nocturnal stools and patient activity. Severe disease activity is classified as a score above 65 [30]. A similar disease activity index exists for CD severity called PCDAI (Paediatric Crohn’s Disease Activity Index). It covers well-being, abdominal pain, number of liquid stools, and abdominal mass and complications, as well as including laboratory results [31, 32].

1.1.6 Treatment strategies

As IBD is a chronic disease, treatment focusses on inducing and sustaining remission. There are therapies specific to these two goals, and not every treatment is suitable for both CD and UC. The clinical classification systems in Section 1.1.5 aid decisions regarding treatment strategy. Treatment often follows a step-up approach, where aggressive therapies are reserved for disease which is resistant to the induction of remission. A summary of available treatments is given in

Figure 2. There are further challenges in treatment of paediatric, and more specifically for younger IBD patients. Height, weight, and body mass index must be monitored, particularly in CD patients, as well as managing puberty and educational needs.

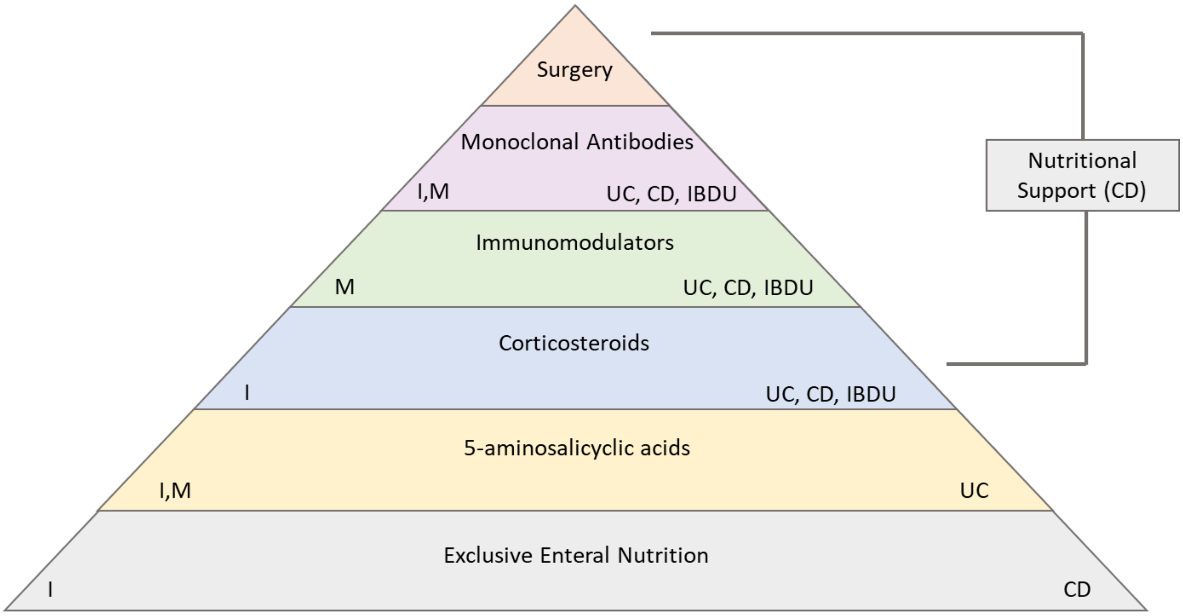


Figure 2 The types of treatment for IBD patients, with the subtype these are suitable for given at the bottom right of each section, and the aim of the treatment in the bottom left (I=induction, M=maintenance). Treatment is usually escalated up the pyramid, starting with safer therapies for milder disease. Sometimes top-down treatment is recommended (starting with monoclonal antibodies), which is a more aggressive treatment strategy that may benefit some patients. Nutritional support for paediatric CD may be given throughout treatment, to address malnutrition and ensure normal growth [33].

1.1.6.1 Induction of remission

Induction therapies are primarily targeted at reducing inflammation for patients. Exclusive enteral nutrition involves the use of a completely liquid diet and is recommended as an initial treatment for CD [33]. It is not recommended for UC patients [33]. Corticosteroids are a treatment appropriate for moderate or severe UC and CD cases [33]. Steroidal treatments may be administered intravenously for acute, severe colitis cases [33]. Due to many side effects when corticosteroids are used long term, it is not recommended for maintaining remission [33]. Corticosteroids can be combined with 5-aminosalicyclic acid medications, and are recommended as an initial treatment for mild to moderate UC [33].

Monoclonal therapies are usually reserved for patients who have disease resistant to remission. These anti-tumour necrosis factor (anti-TNF) treatments, are suitable for both CD and UC patients

[33]. These monoclonal therapies are also known as biologics, and have revolutionised treatment, demonstrating efficacy in adults in the ACCENT [34] and PRECISE [35] trials, and in the REACH [36] trial (88% response rate, paediatric CD). There is some evidence in adults that the introduction of these therapies sooner i.e. a top-down approach starting with the most aggressive treatments, may yield better results [37]. The benefits and risk of this approach have to be weighed carefully, and is currently only recommended in paediatric patients for individuals who have CD with active perianal fistulising disease [38].

1.1.6.2 Maintaining remission

The ultimate aims of maintenance therapies are twofold: clinical remission in order that patients experience no symptoms of IBD, and endoscopic remission, where no inflammation can be seen in the gastrointestinal tract and mucosal healing can be achieved and sustained [8]. As well as being used as an induction therapy, 5-aminosalicyclic acids are used as a maintenance therapy in mild or moderate UC [33]. For CD, the first recommended maintenance treatment is immunomodulators, and this is also suitable for UC patients unresponsive to 5-aminosalicyclic acids [33]. Immunomodulators are slow-acting, so they may be started concordantly with remission-inducing therapies [39]. In patients where immunomodulators are ineffective, biologics are prescribed. Current evidence demonstrates that these drugs can lead to mucosal healing and prolonged remission in paediatric patients [40]. However, these treatments are not an option for every patient, as it is estimated that up to 30% of patients are non-responsive to anti-TNF therapies, and up to 40% may lose responsiveness over time [41]. Small molecule drugs are emerging as possible alternatives that modify specific pathways, such as Janus kinase (JAK) inhibitors. There are promising results in adult trials [41], but currently little evidence of their effective application in paediatric onset IBD. These types of treatments target the underlying molecular cause of an individual’s IBD. A range of small molecule drugs could lead to tailored treatments and less time spent administering therapies that patients will be non-responsive to. However, for this approach to be effective, investigation into the molecular profile of individual patients will need to become the standard.

1.1.6.3 Surgery

Patients with CD can develop complicated disease behaviour, which includes the presence of strictures, or narrowing, in the intestinal tract, and fistulas, where a connection forms between two organs [20]. Relatively common is the perianal fistula, a connection between the anus and skin. The 10-year risk of intestinal resection is reported to be 35.6% in paediatric CD [42]. Other studies have reported surgery rates of 18-35% after 5-year follow-up in cohorts of paediatric and adult-onset CD [43]. Overall surgery rates have declined in CD and UC patients over the past six decades [44], potentially due to the introduction of disease modifying biologics and other more aggressive therapies, and a reduction in time to diagnosis [43]. For some UC patients, a colectomy could completely resolve symptoms. However, there is a risk of complications such as intestinal obstruction, either immediately or further along the disease course that may require surgery [45]. Rates of colectomy are estimated at 10% after 5 years in several cohorts [43].

1.1.7 Long-term risks

A diagnosis of IBD carries an increased risk of developing some cancers, in particular colorectal cancer, lymphoma, non-melanoma skin cancer [42]. Risk factors are from both the inflammation caused by IBD, and the therapies used to treat it. For IBD patients there also appears to be an increased risk of small bowel cancer [46]. The relative risk is highest for specific patient presentations. In UC, those with pancolitis are at highest risk. For CD paediatric onset, ileal disease and stricturing complications in patients increased the relative risk [46]. Some patients will experience extraintestinal complications, once again either caused by the disease or the treatment. This could include complications involving the nervous system, lungs, eyes and bones [47]. As there are complications related to specific treatments, discontinuing therapy during a protracted period of remission is trialled when appropriate. However, this can often result in disease relapse [40].

1.1.8 Age of onset

The presentation of IBD and its prognosis can vary depending on the age of onset. In general, infantile onset is defined as diagnosis before 2 years of age [48, 49]. The presentation of infantile IBD is not usually aligned with either subtype; research found that 71% of a 62 infant patient cohort (<12 months) were diagnosed with IBDU. Of this cohort, 31% required extensive immunosuppression, and 29% were given haematopoietic stem-cell transplantation [50]. Stem cell transplants have been shown improve colitis and gastrointestinal fistulas in those with IL-10 signalling defects, and immunedysregulation polyendocrinopathy enteropathy X linked (IPEX) syndrome [48]. Presentations of these specific immune deficiencies are IBD-like. IL-10 signalling defects manifested as refractory colitis with perianal disease in 100% of patients in a small cohort, with abscesses, perianal fistulas and folliculitis being common [51]. IPEX syndrome patients frequently have watery diarrhoea and enteropathy, in combination with extraintestinal manifestations including type 1 diabetes, neurological and skin conditions [52]. Infantile onset IBD falls within the group of very early onset IBD (VEOIBD), defined as presenting before 6 years of age. The patient group diagnosed before this age are considered to be enriched for IBD caused by a single gene (monogenic), rather than the complex disease of those diagnosed later in childhood, or adulthood [49]. They often have mutations in genes associated with primary immunodeficiencies [49]. VEOIBD is thought to be more difficult to manage, and surgery rates in this population supports this [49]. A diagnosis of UC is more common in VEOIBD, and CD is more common in early onset (<10 years) and paediatric (<17 years) IBD [40].

When considering how age of onset can affect disease management more generally, it should be noted that although an IBD diagnosis during childhood can result in differences in disease course in comparison to adults, the histological, endoscopic and clinical features utilised in diagnosis and monitoring of remission are the same for all age groups [20]. In clinical settings paediatric and adult IBD are not considered to be distinct diseases, aside from in the cases of VEO IBD that require specific treatment protocols such as haematopoietic stem-cell transplantation. The differences between paediatric and adult IBD lie in disease location likelihood, growth impairment and reduced bone density [50], and different tendencies towards severe disease, and consequently surgery and further complications [20]. The age at diagnosis for the Southampton Genetics of IBD study cohort used in the analysis of future is visualised in Figure 3.

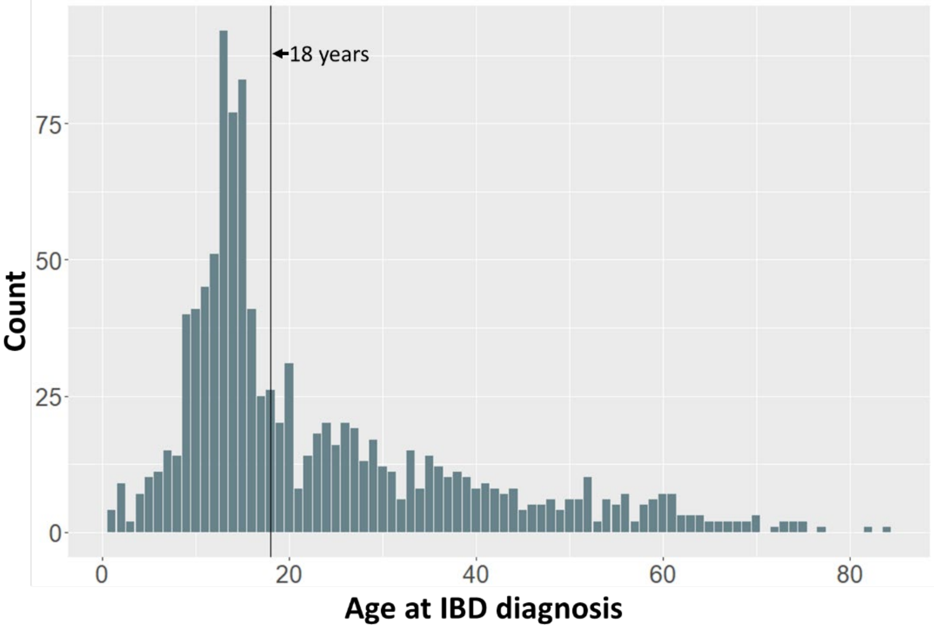


Figure 3 Age at IBD diagnosis for all individuals recruited through the Genetics of IBD study. Vertical line included at 18 years to clearly show the proportion of paediatric and adult onset individuals.

Aside from the differences in common disease locations discussed in Sections 1.1.2 and 1.1.3, there are other differences in presentation and management between adult and paediatric patients. Reports suggest that rates of extraintestinal manifestations are higher in paediatric patients than in adults [20, 53]. Paediatric-onset UC is consistently associated with more aggressive and extensive disease when assessed through endoscopy and histology [54]. This is exemplified by a higher percentage of emergency admittance for acute severe colitis within 5 years of diagnosis for paediatric UC patients [54]. In addition, a study by Van Limbergen et al. of a cohort over 10 years found a 40% colectomy surgery rate in paediatric UC patients, over double that of the adult patients [19]. In general, there is less of a stark difference in disease extent and severity when comparing paediatric and adult CD cohorts. For example, analysis suggests that prevalence of a stricturing endotype, and the development of fistulas after 5 years is similar in paediatric and adult populations [19]. However, there are several clinical presentations indicative of poor CD outcomes for paediatric patients. These include: growth impairment, stricturing endotype and penetrating disease at onset, and severe perianal disease [20]. One study did find significantly more perianal disease (which includes the presence of skin tags, sentinel piles or fistulas) at onset in their paediatric cohort, when compared to the adult IBD patients [53]. A higher percentage of a paediatric IBD cohort was also found to have required a change of drug regime to anti-TNFα therapy during the follow-up period, which was significantly different from the adult IBD cohort [53]. In addition, this paediatric IBD cohort was found to experience significantly more changes in their drug therapy schedules than the adult IBD cohort [53].

1.2 The genomics of inflammatory bowel disease

In Section 1.1, the heterogeneity of inflammatory bowel disease has been described. Patients are diagnosed with a specific subtype, and then according to a sub-classification (CD) or extent and severity (UC), and disease activity will vary for every patient. Many therapies are available depending on the disease course. Successful management is dependent on a prompt and accurate diagnosis, followed by effective treatment. Just as there is clinical heterogeneity, genetic heterogeneity is also present in IBD. Genetics is thought to make a higher contribution to the aetiology of IBD in paediatric patients than in adults [55, 56]. As age of onset increases, the genetic burden changes. In the pathology of some individual's IBD, a genetic profile can be established as an underlying cause of their disease. In many case of infantile or VEOIBD a single gene can be identified as driving disease manifestation (i.e. monogenic IBD) [57, 58]. In early onset IBD, patients may be considered to have a digenic or oligogenic condition, where a small set of genes, potentially from the same pathway, are disease causing. Patients diagnosed later on in childhood or in adulthood are more likely to have a truly complex condition, where many mutations in the genome interacting with environmental components contribute to an IBD phenotype (Figure 4). Some of the clinical heterogeneity is driven by the variation in genes involved. However, it is not clear exactly how the genomics relates to specific phenotypes, and to what extent genomics drives them. It is therefore crucial to understand the genomic landscape of inflammatory bowel disease. As technologies have improved, from early linkage studies, to genome wide association studies, and next generation sequencing, it has become easier to understand and analyse genomics.

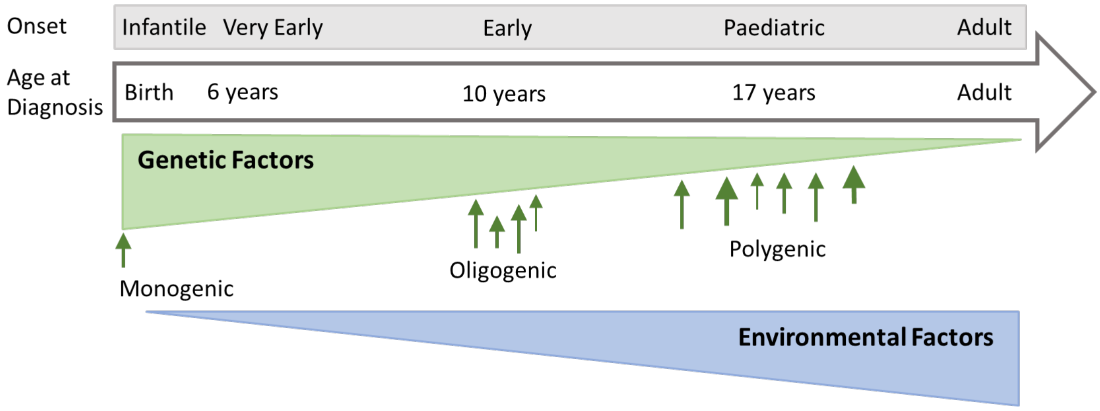


Figure 4 Contributions of genetic and the environmental factors vary depending on age of onset. At a younger age, a single or few genes are likely to contribute significantly to IBD onset. As age increases, environmental factors have a larger impact and many genetic mutations make small contributions to aetiology.

1.2.1 Early discoveries: families, twins and linkage analysis

Early research into IBD genetics focussed on family pedigrees and twin studies. In 1988, the first twin study showed a higher concordance in monozygotic twins with CD (58.3% concordance) and UC (6.3% concordance) than in dizygotic twins (3.6% concordance for twins with CD, 0% for UC) [59]. These two trends in concordance in twins, a higher concordance in monozygotic twins and higher concordance in those with CD, have been confirmed with further twin studies [60]. In 1991, one of the first population-based studies it was observed that in comparison to general population controls, those with a first degree relative that had IBD had a 10-fold increased risk of developing IBD. This was true for both CD and UC patients [61]. This confirms the heritability of IBD, and a higher heritability in CD patients.

Linkage studies were implemented to further the understanding of IBD genetics. Briefly, linkage studies rely on alleles that are located close together on a chromosome being inherited together, as they are highly unlikely to be separated during meiosis (this is also known as alleles being in linkage disequilibrium). Families with individuals that have the trait intended to be studied are genotyped for genetically informative markers. If a marker is close enough to a gene that confers susceptibility to the trait, that genotyped marker will be inherited by those with the trait, allowing researchers to identify regions of chromosomes where causal genes may reside. In IBD linkage studies, a total of nine regions were reported as conferring susceptibility on chromosomes 1, 3, 5, 6, 7, 12, 14, 16 and 19 [1]. Of these linkage studies, the most replicated result was that of the IBD1 locus on chromosome 16 being associated specifically with CD, and not UC. It was later confirmed that the association was with the *NOD2* gene when specific mutations were identified [2, 3]. *NOD2* (Nucleotide-binding oligomerization domain-containing protein 2) is a gene involved in the recognition of bacterial lipopolysaccharides, and triggers an immune response via the activation of NF-κB. Many of the loci identified during this time pointed towards the involvement of genes that could result in a dysregulated immune system if mutated. The IBD3 locus on chromosome 6 contained the Major Histocompatibility Complex (MHC), a region involved in the recognition of antigens, and the IBD5 locus on chromosome 5 contains genes coding for a number of immunoregulatory cytokines [1].

1.2.2 Genome Wide Association Studies

GWAS, or genome-wide association studies, have changed the understanding of the genomics of many complex diseases. Early association studies tested a small number of genes in a modestly sized case and control cohort (unlike linkage studies, individuals in the cohort were unrelated). An allele in a gene was said to be associated with the phenotype being investigated if it occurred at a significantly different frequency when the cases were compared to the controls. In contrast to association-based methods and linkage analyses of a family with a small numbers of markers (300-5,000), GWAS probed across the whole genome. In GWAS, sizable cohorts of cases and controls are genotyped at genetic markers known as SNPs (single nucleotide polymorphisms), with up to half a million SNPs genotyped for each person using commercially available arrays [62]. The frequency of the genotypes among cases and controls are compared, to see if a statistically significant association between any SNPs and disease can be identified. These studies are particularly suited to identifying disease-associated SNPs that are relatively common, and only have a modest effect size. The effect of rare variation is not able to be detected this way, as this would require huge numbers of participants. Additionally, not every associated SNP can be linked with a single gene, if there are a multiple genes in the SNP region or conversely, no genes [62].

IBD was a big winner in the GWAS era, and to date over 230 loci have been identified as associated with IBD [63-65]. The majority of these loci are not linked to a specific subtype, there are 71 exclusive to CD and 30 exclusive to UC [4, 66]. In fact, not only are many loci associated generally to IBD, but approximately 70% are also associated with other diseases that have underlying autoimmunity or immunodeficiency [4]. These loci are involved in pathways with many different functions, including: microbial defence, innate and adaptive immunity regulation, reactive oxygen species generation, autophagy and epithelial recovery [67]. A full review of all genes associated to IBD uncovered via GWAS is outside the scope of this thesis, but in Section 1.2.6 some of the key genes and pathways that have been discovered so far are discussed. Although GWAS studies have found a large number of SNPs associated with IBD, their findings are estimated to account for approximately 37% of the genetic heritability in CD, and 27% in UC [68]. There is potential to uncover some of this missing heritability by identifying variation associated with CD and UC that is rare or private to individuals. This can be achieved by through genetic sequencing.

1.2.3 Variation in the human genome

The human genome contains 3.2×10^9 nucleotide bases, of which there are four types: adenine (A), guanine (G), cytosine (C) and thymine (T). Deoxyribonucleic acid (DNA) is a double-stranded helix organised into chromosomes, and within them regions of coding and noncoding sequence. Only ~1.2% of DNA codes for genes, the sections of sequence that encode the instructions for synthesising proteins. Genes have coding and noncoding regions called exons and introns, respectively. The sequences between genes are called the intergenic regions. The remaining 98.8% of DNA is still of import, as it performs regulatory functions [69].

Each gene has several exons and introns. In order to synthesise proteins from the gene sequence, it is transcribed into messenger ribonucleic acid (mRNA). Then, splicing begins, where the intronic regions are removed from the sequence, leaving contiguous exons in a mature mRNA. Ribosomes translate the modified mRNA into protein, where three bases called a codon code for one amino acid in the protein's structure. Start and stop codons guide the beginning and end of translation by the ribosomes. As the same amino acid can be coded for by more than one codon, there is an amount of redundancy in the genetic sequence. Therefore, not all variation in the human genome will cause disease. Furthermore, different types of variation can have different consequences downstream.

Variation in the genome can be large or small scale. Large scale, or structural, variation refers to copy number variants that cause changes in the sequence longer than 1 kilobase [70]. Small scale variation involves a single nucleotide variant (SNV), or a small number of nucleotides. The main types of small variation are as follows:

- Synonymous SNV: one base in the codon is changed, but this does not change the downstream amino acid.
- Non-synonymous SNV: one base in the codon is changed, and this changes the downstream amino acid.
- Stop-gain SNV: a base change converts a codon that codes for an amino acid, to one that codes for the termination of protein synthesis. This can occur anywhere upstream of the initial stop codon.

- Stop-loss SNV: a base change converts a stop codon to one that codes for an amino acid, causing protein synthesis to continue.

- Indel: The insertion or deletion of a small number of bases. Insertions and deletions that are multiples of three will not affect downstream amino acids (non-frameshift variant). The insertion or deletion of other numbers of bases affects all downstream amino acids as it shifts the whole sequence (frameshift variant).

- Splicing variant: a base change in a region of the sequence that instructs the splicing of the gene's exons and introns.

Humans carry two sets of chromosomes (diploid), one set of maternal chromosomes, and one set of paternal chromosomes. A variant can appear on one or both copies of a gene. This is called a heterozygous genotype or homozygous genotype, respectively. Sometimes there can be a different variant on each copy of the gene, and this is called compound heterozygosity. When a male individual has a variant on their only X chromosome, this is called hemizygosity. When trying to determine disease causal variation, it is important to consider the variant genotype as it relates to the disease inheritance pattern. A dominant inheritance pattern means an individual needs only one copy of a gene with the disease causing variation to inherit the disease, and a recessive inheritance pattern requires both gene copies to be affected to cause disease.

1.2.4 Sanger sequencing

This first generation sequencing method was developed in 1977. This method amplifies a sequence using the chain termination method. Four experiments are performed in parallel, with each experiment containing a single strand of the DNA to be amplified, primers to initiate the synthesis of DNA fragments, and deoxynucleotide triphosphates (dNTPs) for all four bases. In each experiment, dideoxynucleotides triphosphates (ddNTPs) for only one of the four bases are included. When a ddNTP is added instead of a dNTP, which will occur by chance during the reaction, the DNA fragment is terminated. These fragments of varying length are then separated using polyacrylamide gel electrophoresis with a lane for each of the four bases, and the resulting sequence can be read off the gel [71]. Contemporary Sanger sequencing uses ddNTPs that have been tagged with a fluorescent marker specific for each base, therefore only one experiment is performed. The DNA fragments are separated by capillary electrophoresis, and the fluorescent intensity for each base of the sequence can now be read by software [71]. This modern equivalent of Sanger sequencing is still in use in clinical settings, often to validate variants in the genome found during analysis of high throughput sequencing data.

1.2.5 Next Generation Sequencing

The development of second generation sequencing, also known as high throughput sequencing or next generation sequencing (NGS), led to genetic data being generated at increasingly rapid rates for lower costs. In 2001 it cost $100,000,000 to assemble the first human genome, but 20 years later to sequence an individual's genome costs approximately $1,000 [72] (Figure 5). NGS is also called short-read sequencing, as the method involves fragmenting the DNA into segments that could range from 50 base pairs to 600, depending on the method. The sequences generated as part of NGS are known as reads, and correspond to all, or part of the fragments of DNA.

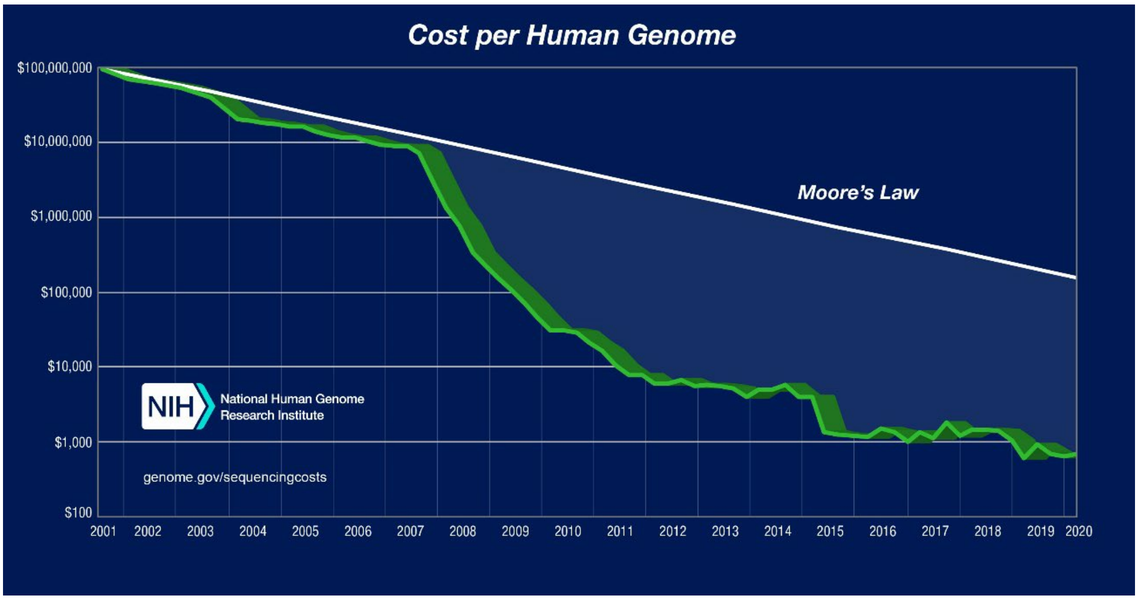


Figure 5 Decreasing costs of human genome sequencing, compared against Moore's Law. Moore's Law is the observation that computing power doubles every two years, but cost decreases [72].

1.2.5.1 Genome, exome and targeted sequencing

There are three main approaches to sequencing an individual: whole genome sequencing, whole exome sequencing and targeted sequencing. In whole genome sequencing, a large volume of data is generated on an individual, as every base is sequenced. This allows the analysis of the protein coding exons, and the introns and intergenic regions that can contain important regulatory sequences that influence transcription and splicing. The amount of data generated is dependent on the depth to which the genome is sequenced. Here, depth of sequencing refers to the number of times each base is included in a sequencing read. A higher depth of sequencing usually gives more confidence to individual variant detection, as more reads containing the same variant mean the variant is more likely to be true, rather than a sequencing error. However, higher depth comes at the cost of bigger file sizes. A genome sequenced at approximately 40x depth will require over 300GB of storage space, as not only do the raw data files need to be stored, but also files that are required for analysis of the genome (three file types need to be stored, fastq, binary alignment map and variant call format file, discussed in Chapter 3) [73]. Generating these files for analyses is computationally and time intensive. The large volume of data generated can also create issues for interpretation, as there are roughly 3.7 million variants in every person's genome [74], and the vast majority will not cause or impact disease. Interpreting variation can be particularly challenging when analysing the intronic and intergenic regions. Current bioinformatic tools are less equipped for non-coding variation, as the connection between mutation and the downstream effects on protein production and pathways can be more obscure.

In comparison to whole genome sequencing, whole exome sequencing produces substantially less data. To generate the same files listed for analysis previously, at approximately the same depth (50x) would require only 13GB of storage (own data). While only 1.2% of the human genome are coding regions [75], it is suggested that up to 85% of disease-causing mutations are contained in the exons [76]. Therefore, using exome sequencing can reduce the computation power, the computational and analysis time, and still uncover disease causing mutations. There are approximately 26,000 variants for each individual's whole exome sequencing data [77], but bioinformatic tools are more equipped to assess the causal nature of these. As only specific sections of the genome are being sequenced, due to the sequencing technique, the depth of sequencing tapers towards the end of each exon (Figure 6A). This can cause difficulty in generating enough data at the ends of exons to confirm variants residing there. Due to this, splicing variants and variants in start and stop codons may be missed. It can also be very challenging to detect copy number variants with whole exome sequencing, as DNA is usually fragmented into smaller pieces during exome sequencing than in genome sequencing. This fragmentation also causes a variable depth in the regions sequenced. Reassembling these large regions of changed sequence is difficult.

Targeted sequencing further reduces the DNA sequenced to specific genes or regions that are to be analysed in an individual. This method is currently applied for diagnosis of a disease that has a well-defined phenotype and an associated panel of genes. Additionally, it can be used to sequence a gene or region of interest very deeply. This approach also reduces data analysis time. However, if the causal mutation is not found in the initial investigation, multiple rounds of sequencing different genes can become more expensive that whole exome sequencing. Targeted sequencing also does not have the advantage of being able to apply different gene panels *in silico*, as in whole genome or whole exome sequencing. With the latter two methods, sequencing data can be revisited when new evidence of disease specific causal mutations comes to light. For the study of IBD genetics, whole exome sequencing is used as the best balance between data produced and interpretability for a longitudinal cohort study of a complex disease.

1.2.5.2 Whole exome sequencing technology

Whole exome sequencing (WES) consists of four steps: 1) Library preparation; 2) Amplification; 3) Sequencing (Figure 7); and 4) Data Analysis. During library preparation for whole exome sequencing the extracted DNA to be sequenced is sheared randomly into fragments, either using a mechanical method such as ultrasonication shearing and nebulisation, or enzymatic digestion (biological method) [78]. Adaptors are ligated on either end of the produced library fragments. Tag sequences are ligated onto one or both fragment ends, depending on whether single-end sequencing or paired-end sequencing will take place. During paired-end sequencing, both forward and reverse strands are sequenced. By knowing the total DNA fragment length and the length of the forward and reverse reads, the distance between the reads is also known (the inner distance) (Figure 6B). This makes the process of mapping the reads to the reference sequence more accurate and efficient. After adaptors have been attached, DNA or RNA baits, along with oligonucleotides that are complementary to the adaptors, are added. These baits hybridise to the DNA fragments that are exonic. This allows fragments that are not the target for sequencing (non-exonic regions) to be washed away, and sequences that are the target to be pulled down for sequencing [79]. Targeted sequencing is enabled by capture kits that are designed so that the correct sequences are pulled down. In this case the target is the whole exome. These sequences will be amplified in the next step.

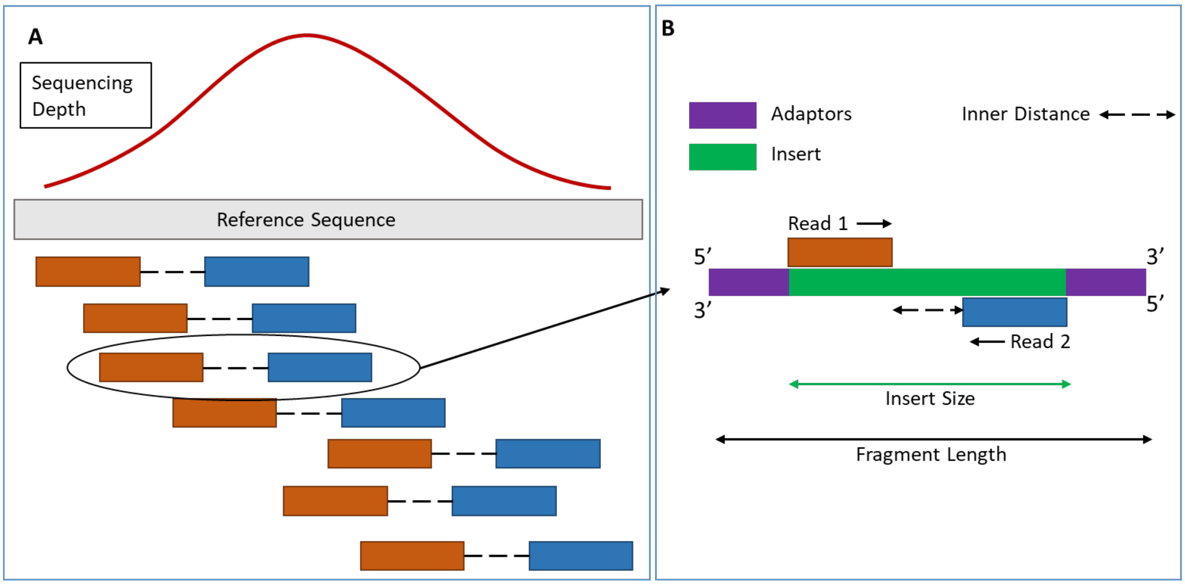


Figure 6 Paired end sequencing. A) The sequencing of fragments is achieved through a forward strand read (read 1) and a reverse strand read (read 2). In whole exome sequencing, the introns are not sequenced, so at the points where the exon ends there is a tapering of reads due to their stepped arrangement. There are fewer reads to overlap to get an increased sequencing depth at the ends. B) Reads are sequences corresponding to part of the original fragment. There is a known distance between the two reads called the inner distance. The read length will vary, for whole exome sequencing this would usually be 50-150 base pairs depending on the method used. The fragment lengths will also vary but will invariably be longer than the read lengths (approximately 150-300 base pairs).

Illumina is a biotechnology company that currently dominates the global high-throughput sequencing market. They perform DNA amplification via cluster generation. During cluster generation, the adaptors on the templates hybridise with complimentary oligo primers on the surface of a flow cell. A new strand is synthesised by extending the oligo primer to create a strand complementary to the DNA fragment. Then the bridge amplification technique is used, where the unconnected end of the newly synthesised strand hybridises to another oligo primer, and another strand is extended from this primer. Once the two strands are denatured from each other, a forward and reverse strand has been synthesised and are attached to primers on the flow cell. Clusters are generated by repeating the bridge amplification process thousands of times. Finally, the reverse strands that were generated are cleaved off, leaving the forward strands to be sequenced.

The most common method for step three is sequencing by synthesis. This method consists of a series of cycles of 1) Incorporation; 2) Imaging; and 3) Cleavage. A mix of the four fluorescently tagged nucleotides and DNA polymerase are added to the reaction. One base at a time hybridises to the DNA templates synthesised during cluster generation, as there is a reversible terminator on each nucleotide. During imaging the florescent tags are excited, for example by a laser, and the emission spectra is captured. In the final step of the cycle the fluorescent tag and reversible terminator are cleaved, allowing a new nucleotide to be added for the next cycle. If the sequencing is paired end sequencing, strands generating during sequencing are denatured and washed away. Another round of cluster generation follows, where the forwards strands are cleaved leaving the reverse strands to be sequenced.

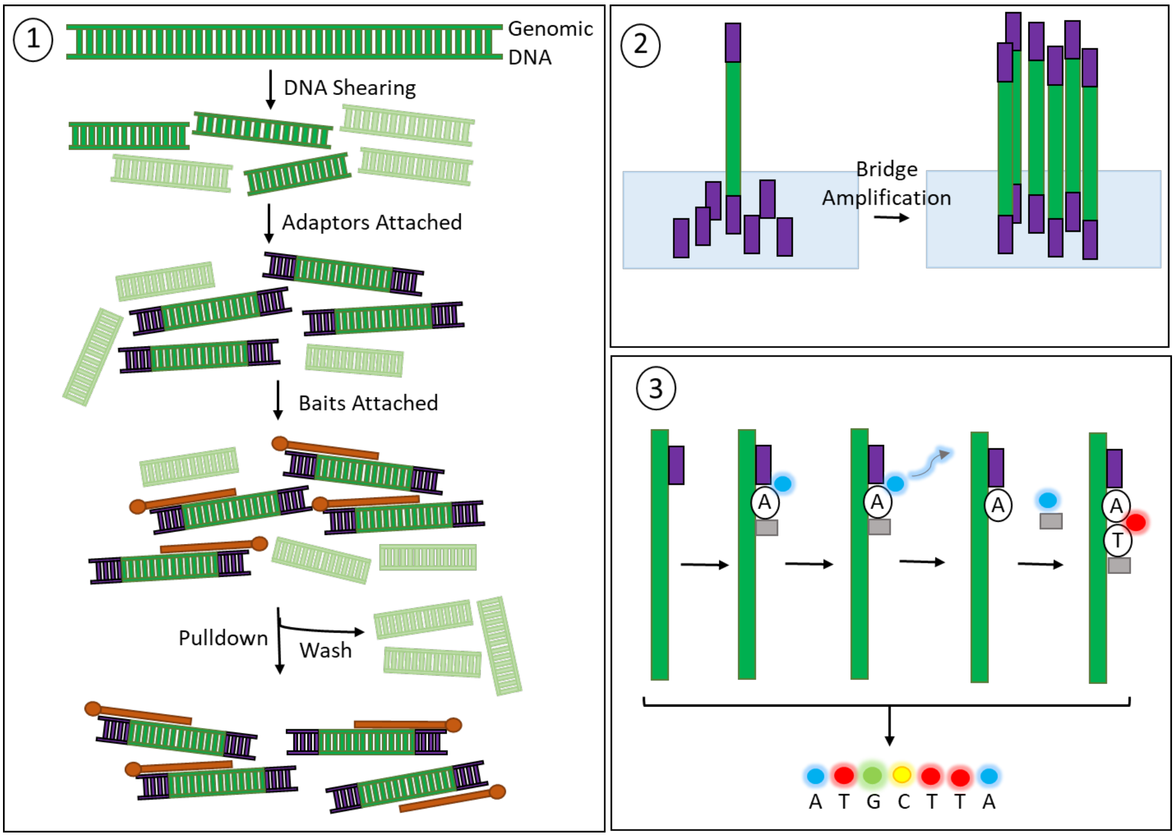


Figure 7 Schematic of the whole exome sequencing process 1) Library Preparation: the process of shearing DNA into fragments and pulling down fragments that are targets for sequencing. 2) Amplification: this process is as described by Illumina where thousands of copies of a DNA fragment are synthesised in order to amplify the signal during the next step. 3) Sequencing: the sequencing by synthesis method where nucleotides with a fluorescent marker and reversible terminator are added one by one. The fluorescent marker is excited and emits a frequency specific to the base. The marker and terminator are cleaved and the process repeats, giving the sequence.

1.2.5.2.1 Analysing whole exome sequencing data

Data from whole exome sequencing is commonly output in the form of a fastq file containing quality information alongside each sequencing read. The first step is alignment to the reference genome using bioinformatic tools such as the Burrows-Wheeler Aligner [80]. The reference genome has changed over the years since the first assembly during the Human Genome Project, with each iteration filling in gaps in the human reference. The current genome build is GRCh38 (hg38) comprised of 11 individual's genomic sequences [81]. The most recent builds also have multiple alternative sequences in regions that are very diverse, particularly among different ancestry backgrounds. After alignment, regions that differ from the reference are identified (variant calling), and finally variants are annotated with useful information for their interpretation (detailed information regarding exome sequencing data processing methods are given in Chapter 3). Many databases and bioinformatic scoring systems exist to annotate variants and help identify probable causal mutations. GnomAD [82] and the 1000 genomes project [83] databases provide information on variant frequencies in a population. PolyPhen-2 [84] and SIFT [85] both score the likelihood that nonsynonymous variants are damaging, dbNSFP [86] scores nonsynonymous and splicing variants, and GERP [87] scores how conserved each SNV is likely to be. CADD [88] and DANN [89] score variants' deleteriousness using machine learning and deep learning, respectively. Both CADD and DANN can score all types of small variation, including synonymous SNVs and indels. The databases ClinVar [90] and HGMD [91] collate information from literature regarding potential variant pathogenicity, with the former relating this data to recorded clinical phenotypes.

Once variants have been annotated, the variant frequencies and type are considered. Variants that are not rare (>1% population frequency) can be filtered using information from databases such as gnomAD [82], and synonymous variants excluded. The next steps depend on the purpose of the variant analysis. If the analysis is being performed for clinical diagnostics, then it is likely that deep phenotyping has been established. Additionally, a family history may be available so that only variants that conform to the likely inheritance patterns will be considered. Literature searches can identify a shortlist of genes, as well as the use of Genomics England's PanelApp, which stores virtual gene panels [92]. ClinVar [90], and literature searches can help identify likely pathogenic variants. If the variant analysis is being performed for research, then the standard candidate gene list may have already been exhausted. For Mendelian diseases, the American College of Medical Genetics (ACMG) guidelines can be used to interpret variants of unknown significance [93]. In all cases, the aforementioned bioinformatic scoring tools would be used to determine a likely causal variant. However, any variant identified this way will need to be functionally validated to discover if it does change downstream mechanisms as suggested by the variant analysis. A summary of the strategy for analysing sequencing data to find causal variants is given in Figure 8.

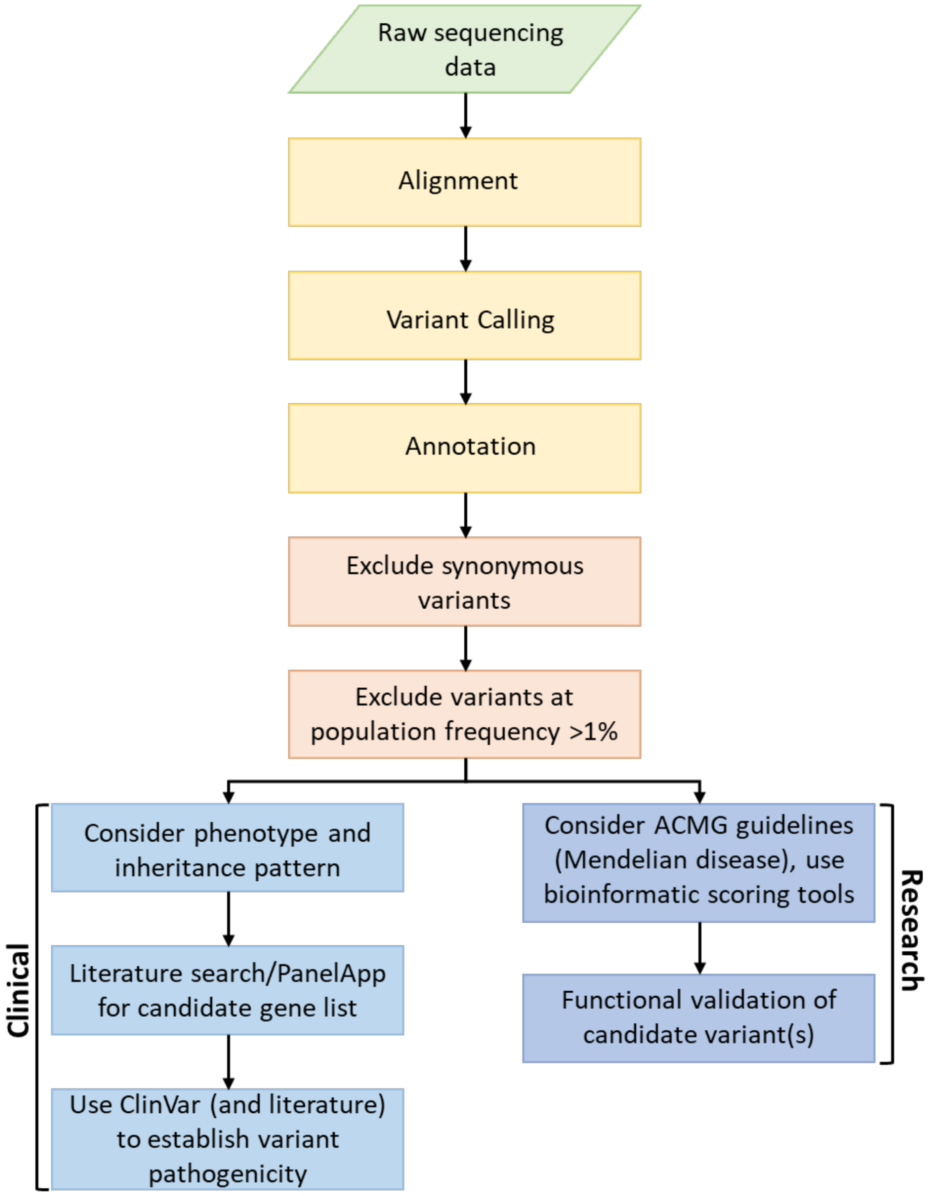

Raw sequencing data
Alignment
Variant Calling
Annotation
Exclude synonymous variants
Exclude variants at population frequency >1%
Consider phenotype and inheritance pattern
Literature search/PanelApp for candidate gene list
Use ClinVar (and literature) to establish variant pathogenicity
Clinical
Consider ACMG guidelines (Mendelian disease), use bioinformatic scoring tools
Functional validation of candidate variant(s)
Research

Figure 8 Workflow for processing and analysing sequencing data. Simplified steps are 1) Data processing; 2) Variant filtering; 3) Evaluating remaining variants to find probable causal variants. This third step varies depending on whether the analysis is for clinical diagnostics, or occurring in a research environment.

All filtering strategies have limitations. Firstly, synonymous variants can be disease causal, in particular there is evidence that these variants can impact splicing. The appropriate splicing of exons is contingent on specific sections of the exonic sequence signalling to splicing machinery [94]. As this process is not fully characterised, some variants that will impact splicing are mislabelled as synonymous. This illustrates a key constraint, that the assessment of variation is only as good as the bioinformatic tools available. This limitation can be partially mitigated by consulting many different databases to obtain a consensus view of whether a variant is likely to be damaging. Additionally, compound heterozygotes can be difficult to identify with whole exome sequencing data. It cannot be confirmed whether two different variants in the same gene have impacted one, or both copies of the gene. There are also multiple transcripts available for each gene, and each transcript will place the variant in a different position. These transcripts are supported by different levels of evidence, and a well-defined gene may have more than one transcript supported by robust evidence. Multiple transcripts make analysing variants more challenging. Many pipelines are not equipped for cases where variants will have different impacts on the structure and function of proteins depending on the transcript used. In cases where pipelines are equipped, strategies for prioritising different, potentially all well-evidenced, transcripts must be developed [95]. This adds to the already time-intensive task of analysing whole exome sequencing data.

1.2.6 Genes and pathways of inflammatory bowel disease

Genes discovered through GWAS allowed further elucidation of pathways involved in the pathology of IBD (Figure 9). These pathways are interlinked, with many proteins contributing to the activation of multiple downstream mechanisms. High throughput sequencing enabled researchers to gain granularity regarding specific variation within these genes at scale. Further, NGS technology has led to an increase in the identification of rare variants specifically associated with earlier-onset IBD, either in novel genes, or in those that were already known to be associated with the disease.

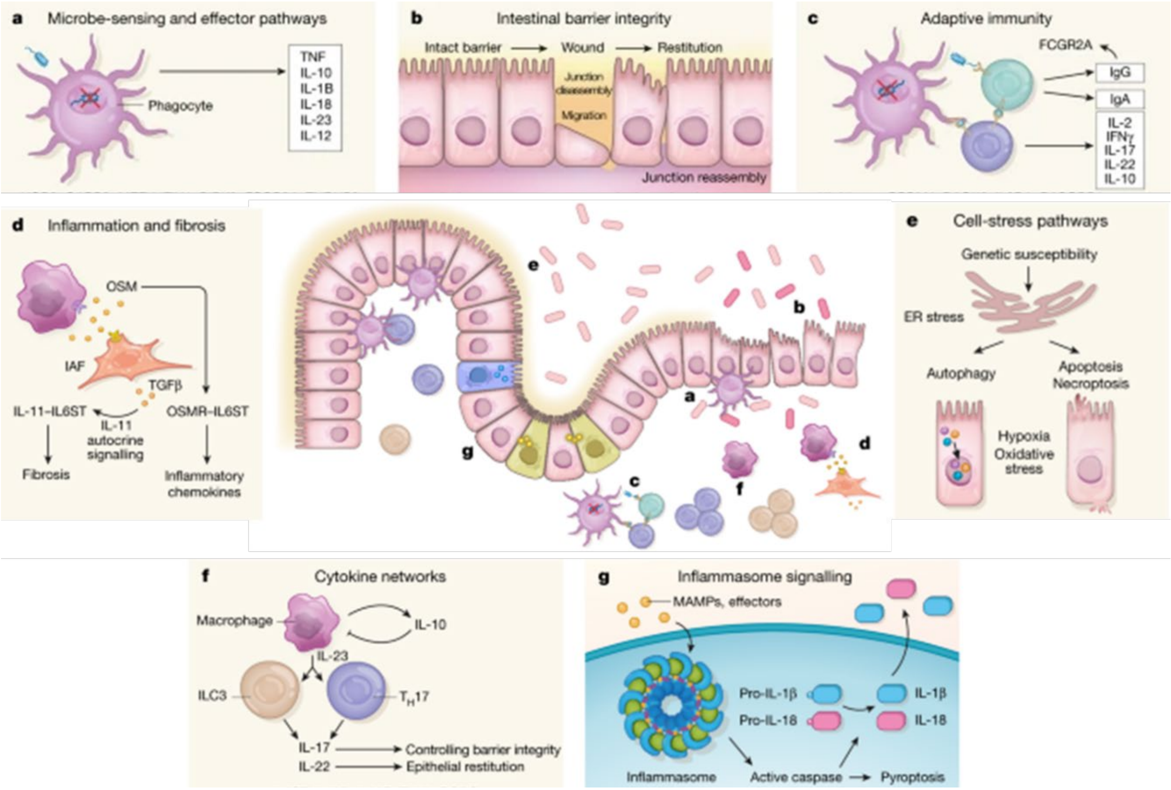


Figure 9 Pathways and mechanisms potentially affected by IBD pathogenesis. a) Microbial sensing by phagocytes that triggers pro-inflammatory cytokines; b) Intestinal epithelial barrier regulates contact between immune cells, and microbes and antigens; c) Adaptive immunity, where T-cells and B-cells facilitate immune response; d) Inflammation and fibrosis. e) Cell stress processes causing autophagy and apoptosis; f) Cytokine networks; g) Microbial recognition by inflammasome complexes. Image adapted from [96].

1.2.6.1 Microbial recognition

The mechanisms that recognise microbial antigens are highly implicated in CD. Immune cells such as macrophages and dendritic cells, as well as cells in the intestinal epithelium have the ability to sense and recognise these molecules [97]. Pathogen-associated microbial proteins are recognised by different types of pathogen-recognition receptors, including toll-like receptors (TLRs) and nucleotide-binding oligomerisation domain-like receptors (NLRs) [97]. The NLRs *NOD1* and *NOD2* recognise γ-D-glutamyl-meso-diaminopimelic acid and muramyl dipeptide, respectively [97]. This can prompt the activation of the NF-κB (nuclear factor- κB), MAPK (mitogen-activated protein kinase) and IFN-β (interferon-β) signalling pathways (Figure 10) [97]. NF-κB signals the activation and differentiation of cells involved in the innate and adaptive immune response [98], and IFN-β signalling mobilises macrophages to resolve bacterial inflammation [99]. *NOD2* is the most well defined risk gene in CD [2, 3]. *NOD2* variants could affect the immune response in different ways. Decreased NF-κB signalling caused by impaired *NOD2* could reduce the antimicrobial response, causing a pathogenic microbial invasion [100]. Alternatively, *NOD2* variants could cause decreased inhibition of TLR2, leading to an excessively upregulated response from adaptive immune cells [101]. In addition, a specific insertion (3020insC) has been shown to decrease interleukin-10 (IL-10) expression, a cytokine that can downregulate the immune response [97].

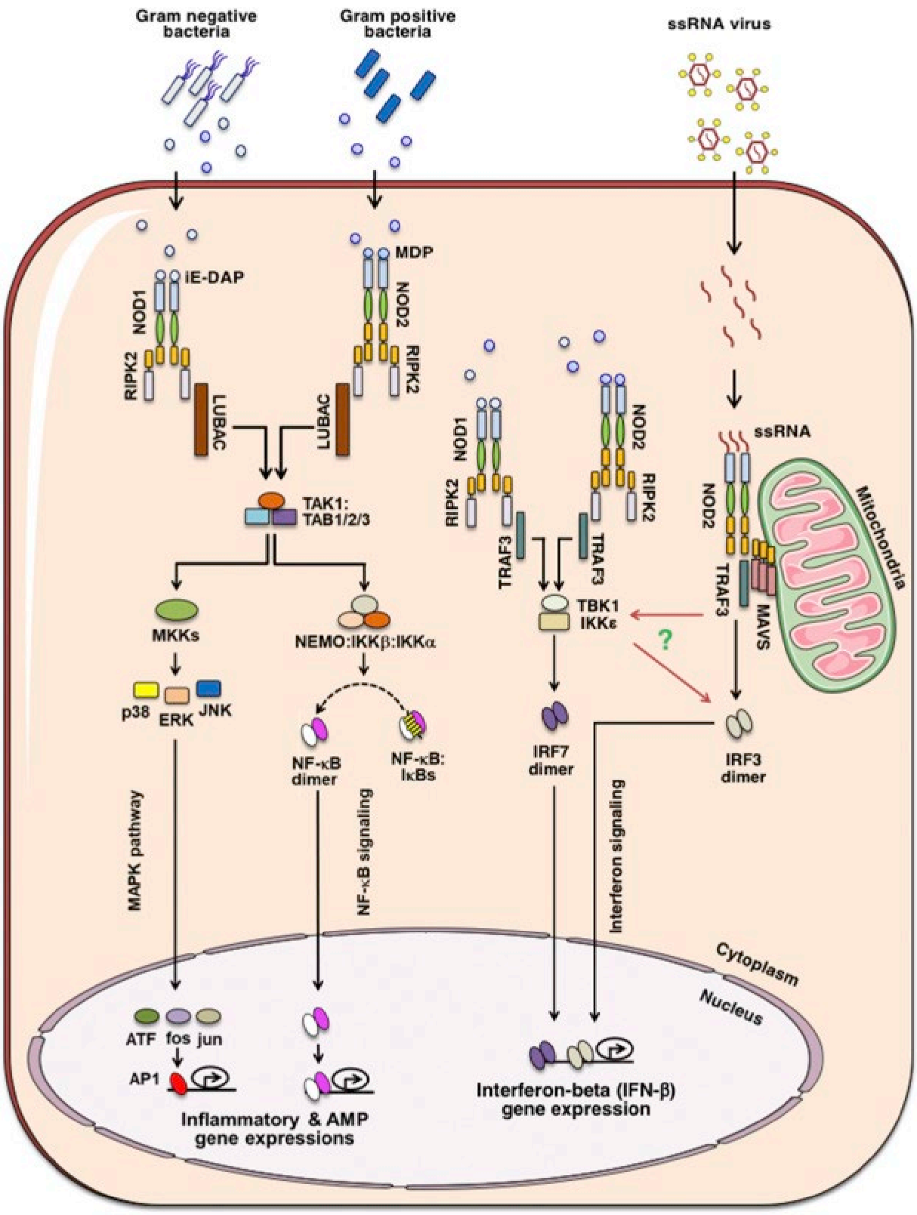


Figure 10 Recognition of bacteria by *NOD1* and *NOD2* proteins triggers the activation of NF-κB, MAPK and IFN-β signalling. Image adapted from [102].

1.2.6.2 Innate immune response

The innate immune response is the first defence against pathogens, and is not specific to any pathogen. Several components of the innate immune response are implicated in IBD, including epithelial barrier function and autophagy [97]. The first layer of defence of the intestinal epithelium is the mucosal layer, and mucin genes are an important component of maintaining this. In CD, abnormal expression of mucin genes in comparison to controls has been observed, with decreased expression of *MUC1* in inflamed ileum, and *MUC3*, *MUC4* and *MUC5B* in uninflamed ileum [97]. The integrity of the epithelial barrier is essential to correctly regulate foreign bodies coming into contact with immune cells. The structure is maintained by tight junctions, adherens junctions and desmosomes. Increased intestinal permeability is a feature of both CD and UC, however many genes from IBD GWAS that code for components or regulators of the epithelial barrier have been specifically linked to UC (*HNF4A*, *CDH1*, *LAMB1* and *GNA12*) [97].

Two genes with a role in autophagy that have been associated with CD are *IRGM* and *ATG16L1 [97]*. Further, *ATG16L1* is thought to be closely linked with *NOD2*; bacterial activation of *NOD2* triggers autophagy, and epithelial and dendritic cells with variants found in CD show antibacterial autophagy defects [67].

1.2.6.3 Reactive Oxygen Species

CYBA, *CYBB*, *NCF1*, *NCF2* and *NCF4* are genes implicated in monogenic IBD-like disease, specifically chronic granulomatous disease (CGD) [103]. This is a primary immunodeficiency that affects the innate immune system, as a significant majority of mutations that cause CGD are loss-of-function mutations that result in an absence or reduction of protein subunits that form the NAPDH oxidase complex [104]. When the NADPH oxidase complex is activated, reactive oxygen species (ROS) are produced as an innate immune response to any combination of the following: microbes, activated pattern recognition receptors, and phagocytosis [103]. It is the failure of this mechanism that leads CGD patients to be very susceptible to infection [104]. However, an overproduction of ROS can activate the generation of pro-inflammatory cytokines such as TNFα (tumour necrosis factor α) [103]. ROS influences pro-inflammatory cytokine production through the NF-κB signalling pathway, a process that can also be induced by *NOD2* [97]. These factors make the genes encoding the proteins of the NADPH complex of interest for elucidation of the genetics of IBD.

There are seven different isoforms of NADPH oxidase (NOX) genes, and in particular NOX1, NOX2, NOX3, Duox1 and Duox2 have been reported as being expressed in part(s) of the gastrointestinal tract [105]. The genes *NOX1*, *NOX2*, *NOX3*, *NOX4*, *NOX5*, *DUOX1*, *DUOX2*, *CYBA*, *RAC1*, *RAC2*, *NOXA1*, *NOXO1*, *RAP1A*, *NCF1*, *NCF2*, *NCF4*, *DUOXA1* and *DUOXA2* code for subunits in one or more of the NADPH oxidase complex isoforms [105]. In *NOX1*, a stop-codon mutation has been identified in an early onset IBD patient [106], and missense and loss of function mutations have been found in VEOIBD patients [107]. Missense mutations in *NOX1* and *DUOX2* have also been found together in a VEOIBD cohort, and associated with Paneth cell metaplasia (specific epithelial cells in areas where they do not normally occur) [108]. A splicing mutation in *CYBA* [109], and missense mutations in *CYBA*, *CYBB*, *NCF1*, *NCF2* and *NCF4* [110], have all been reported specifically in CD patients. A population based study found a significant association between perianal disease behaviour and *NCF4* mutation [111].

1.2.6.4 Adaptive immune response

The adaptive immune response (secondary immune response) targets specific antigens and occurs after the innate immune response. In IBD, dysregulation around adaptive immunity is linked to a loss of homeostasis between T regulatory-cells (T_{reg}-cells, regulate the immune response) and T helper-cells (T_H-cells, involved in activation of a number of immune cells) [67]. It is thought that T_{reg}-cells do not sufficiently control T_H-cell response in IBD [67]. Increased levels of T_H17 cells have been found in CD and UC patients [112]. T_H17 cells are induced by the IL-6, TGF-β and STAT3 expression, and their proliferation stimulated by IL-23 [112]. T_H17 cells subsequently produce the cytokines IL-17A, IL17F, IL21 and IL-22, the majority of which support pro-inflammatory actions [112]. *STAT3*, *IL-23R* (IL-23 receptor) and *JAK2* are some of the genes implicated in T_H17 dysregulation [67]. Traditionally, T_H1 cell responses have been associated with CD, and T_H2 cells with UC [67]. This is due to the levels of different cytokines in the different subgroups: CD patients are reported to have high levels of IL-2 and IFN-γ, and IL-5 and IL-13 are present at high levels in UC patients [112]. B regulatory (B_{reg}) cells are also implicated in IBD [67]. Defects in these cells can result in a failure to upregulate anti-inflammatory cytokine IL-10. GWAS implicated *IL10* as a central immune regulation gene in the understanding of IBD [67].

1.2.6.5 Monogenic inflammatory bowel disease

In cases of VEOIBD, it is more likely that the disease is attributable to one gene [113]. Currently, 94 genes have been reported in the literature and implicated in monogenic IBD, or IBD-like conditions (Supplementary Table 1) [57, 92, 114-142]. Many of these genes are associated with a primary immunodeficiency, or autoimmunity mechanism. Mechanisms affected by these genes include T and B cell production and regulation, phagocyte function and epithelial barrier function. The discovery of these genes is an important step towards personalised medicine for VEOIBD, as in these cases it is possible to pinpoint the exact mechanism by which IBD or IBD-like disease manifests. For example, Mao et al. found that a *CARD8* mutation that causes monogenic CD could not be treated with anti-TNFα therapy due to the impediment of interactions between *NLRP3* and *CARD8*, but that patients were more responsive to IL-1β inhibitors [116]. Further, in a case where a young child presented with intractable IBD, whole exome sequencing analysis revealed a hemizygous mutation in the *XIAP* (X-linked inhibitor of apoptosis) gene. The precise diagnosis meant the patient received an allogeneic haematopoietic progenitor cell transplant, and following the treatment there was no recurrence of gastrointestinal disease [6].

1.2.7 Integrating genomic information

As genomics becomes more commonly used for cancer and rare disease in a clinical setting, research focus has shifted towards using genomic data for more complex situations, leveraging it to predict prognoses, or in diagnosis for complex disease. GWAS data has been used to devise polygenic risk scores, usually a sum of weighted risk loci. Uses for polygenic risk scores include selecting the appropriate treatment for individuals and screening a disease susceptible population [143]. Polygenic risk scores are limited as, although they can be derived from the whole genome or whole exome, they usually select specific variants to contribute to these scores. It also assumes a linear relationship between factors. High-throughput sequencing methods go some way to mitigating these issues. Polygenic risk scores have been constructed for IBD cohorts. Rarely have scores been connected back to a specific clinical phenotype, or the specific genetics driving differences in individual’s polygenic risk scores been thoroughly examined. The majority of studies focus on case vs control studies. Chen et al. synthesised risk scores for CD and UC patients, and controls in a cohort from Australia and New Zealand. In a CD subgroup for which they had phenotypic information, they found statistically significantly higher genomic risk scores in those that required bowel resection, patients with a younger age of onset, and individuals with more ilealic inflammation than colonic [144]. They did not report the genetics driving these differences. Vancamelbeke et al. focused on generating risk scores from a subset of genes involved in intestinal epithelial barrier dysfunction. They report significantly higher scores in CD and UC when each subtype was compared against controls. These differences were driven by *MUC19*, *MUC22*, *TFF1* and *PTGER4* in CD, and *MUC21*, *MUC22*, *GNA12* and *HNF4A* in UC [145]. Another study from Serra et al. conducted analyses on a VEOIBD cohort. They confirm a significantly higher risk score for VEOIBD in comparison to controls. They also confirm a polygenic component to VEOIBD, but were unable to quantify the contribution of common risk variants due to a lack of monogenic diagnoses in the cohort [146].

Understanding of the genomics of IBD has improved dramatically in the past two decades, but there has been limited translation of this knowledge into clinical settings. Some monogenic conditions with IBD-like manifestations are able to be treated specifically, for example stem cell treatments for IPEX patients who harbour a mutation in *FOXP3* [52]. This kind of personalised care must be extended to all IBD patients for better outcomes, and to achieve this there must be an understanding of the links between genotypes and phenotypes. Currently, our ability to generate patient’s genomic data is much faster than the ability to effectively analyse large volumes of data, particularly if the analysis is cohort-based rather than on an individual level. The scale and complexity of high-throughput data requires different tools for interpretation.

1.3 Machine learning

Machine learning sits under the umbrella term of artificial intelligence, along with other intelligent system methodologies. Machine learning involves the implementation of algorithms to perform specific tasks, usually classification or regression. Classification problems can involve two or more groups, for example sorting patients according to the treatment they would be responsive to, or classifying them by disease subtype. Regression problems seek to predict a continuous variable. Examples include predicting the correct treatment dose for patients, and estimating the length of a patient's hospital stay. These algorithms are not instructed, but infer patterns from data that are often obscure and non-linear. All artificial intelligence methods are unlike traditional statistical analyses, because these algorithms are intended for prediction, rather than inference. Machine learning models are built with the intent to focus on one specific problem, or patterns in a particular type of data. This is in contrast to other artificial intelligences, where the aim may be to develop a system that is capable of handling many tasks. Key to the effectiveness of machine learning, particularly if it is to be applied in a medical setting, is the ability of a model to be robust and generalisable to new data.

1.3.1 Supervised and unsupervised learning

In the field of machine learning, there are two main types: supervised and unsupervised learning (Figure 11). During supervised learning, the aim is to train a model to recognise patterns associated with a specific outcome. The training data has a number of variables associated with each individual data point (n). One of these associated variables is identified as the "outcome" variable that machine learning is attempting to predict, and can be a continuous or discrete variable. During machine learning training, the outcome variable is visible to the model, and the patterns in the data that relate to the outcome variable value for each data point are learned. Usually in a supervised machine learning workflow data is unevenly split into training and testing data, where the majority of the data goes towards training the model. The model produced by the training is applied to the testing data, where the outcome variable is hidden and the model predicts the outcome variable for each data point according to the rest of the variables. The success of a machine learning model is measured by its performance on the test data, as to be an effective model it must generalise well to unseen data.

Analysis conducted utilising unsupervised learning is more exploratory. Here, there is no outcome variable to be predicted, and the model is constructed based on all the data. Unsupervised learning methods are based on clustering the data based on the patterns present. Apart from discovering new groupings, these methods are also useful if there is no gold standard available, or if the current ground truth could be considered unreliable. This is a disadvantage of the supervised method, as generalisability of a model could be undermined by an outcome variable that does not accurately reflect the ground truth. However, as unsupervised learning models are free to cluster according to any observed pattern, they may be more sensitive to bias in the data, for example inherent sex differences in patients, or batch effects in data.

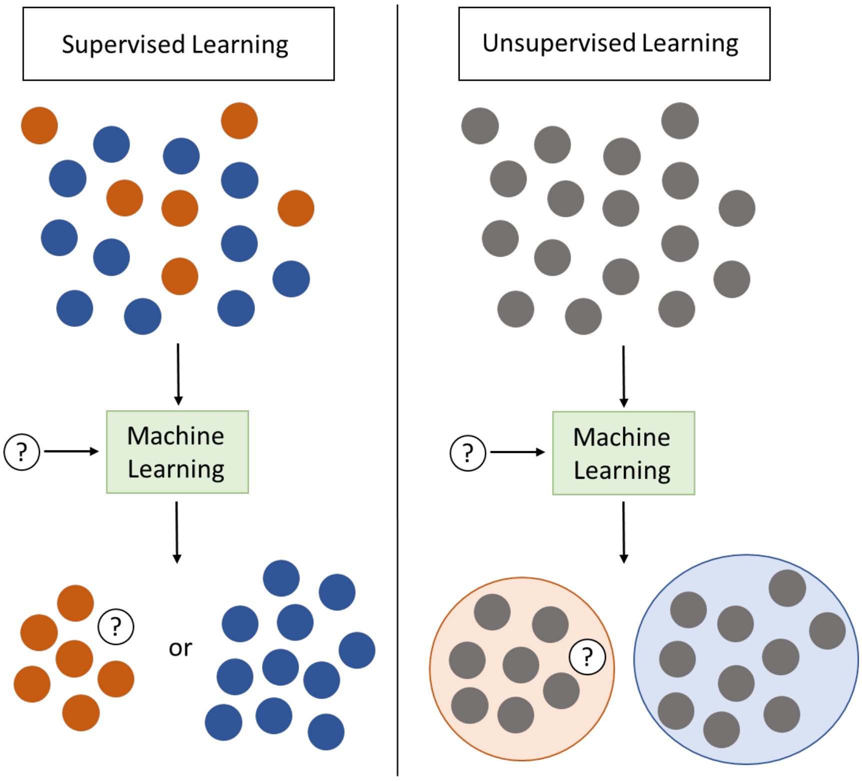


Figure 11 Schematic of the basic principle of supervised and unsupervised approaches. In the supervised method, the groups the data belong to (orange and blue) are known, and the machine learning sorts the data into these groups according to the variables associated with each data point. In the unsupervised approach, groups are not known (grey), and so the learning method clusters the data based on the similarities within. In both cases, the machine learning models can be applied to new unknown data, which is then categorised or clustered accordingly.

There are other types of machine learning. Semi-supervised learning utilises data with and without a labelled outcome variable for training. It can be used when the process to define the outcome variable is laborious. Reinforcement learning features an iterative training process that relies on a feedback loop recording model success and failure in order to better the model performance each time. The review and use of these methods is outside the scope of this work, but further information is supplied elsewhere [147].

1.3.2 Machine learning algorithms

The choice of machine learning algorithm is dependent on three main considerations. First is the overall aim of the machine learning: is the analysis intended to be exploratory, or is there a specific prediction problem? This question will decide whether supervised or unsupervised approaches are the best fit. The second consideration is the data. What is the size of the data and what type of data does it contain? Some methods perform better on smaller or larger data sets. Additionally, some types of data require specific methods, for example to extract and use critical information from free text, the use of natural language processing is key. Another data consideration is whether it is expected to contain linear and/or non-linear relationships, and so whether a linear or nonlinear method will be better suited. Lastly, which is more important: performance or interpretability? As methods become more complicated, the processes that are used to attain a high accuracy are more inscrutable. Even if all considerations are made, it is still often common practice to trial many different models in order to find the best performing one. In some cases model results are combined, where the majority consensus of all models for the prediction for each data point is taken as the final result. Table 1 gives an overview of some of the different models that can be used to illustrate the breadth of methods currently available for machine learning.

Table 1 Description of supervised and unsupervised machine learning methods [148]

Method	**Machine Learning Type**	**Description**
Linear Regression	Supervised	A regression method that attempts to fine a line that will fit the most number of data points in the predictor space.
Logistic Regression	Supervised	A version of a regression model that is instead used for classification.
Neural Networks	Supervised and Unsupervised	A group of methods that are loosely based on the structure of the brain, with a series of variably weighted, nested, nonlinear function that processes data points in order to classify, regress or cluster data.
Random Forest	Supervised	An ensemble method that forms a large number of decision trees. These trees iteratively divide the predictor space through a series of binary questions that will allow data points

		to be sorted according to the outcome variable. Each tree sees a subset of the data and the decisions are aggregated.
k Nearest Neighbours	Supervised	Compares the features of each data point to its (k) nearest neighbours, and positions the data point in its according cluster related to the outcome variable.
Support Vector Machine	Supervised	The method partitions the predictor space into two via a decision boundary. Data points fall on either side of the boundary according to the outcome variable. This can be adapted for regression or multi-class classification problems.
Hierarchical Clustering	Unsupervised	Creates a hierarchy of the number of clusters the data can cluster into in a dendrogram format. The number of clusters selected takes up the most vertical space in the dendrogram.
K means Clustering	Unsupervised	Clusters the data points into the specified number of (K) clusters iteratively until no improvement to the “closeness” of the data points in each cluster can be made.

1.3.3 Feature selection

Increasingly, machine learning involves the use of very large data sets, for example the use of genomic or transcriptomic data sets to understand disease. These types of data sets can be highly dimensional, meaning a large number of variables, or features (f), are associated with each data point (n), such that the total number of features (F) is much larger than the total data N (F >> N). This type of dataset can be very noisy, and include data that may not contribute to the current classification or regression task. Including every feature may obfuscate the signals in the data, leading to a machine learning model that cannot regress or classify according to the aim. For this reason, a number of feature selection methods exist, in order that the maximal amount of information is retained for the machine learning task, with the smallest possible number of features.

One of the simplest ways to reduce the dimensionality of the dataset is to remove features where the variance is zero, or very low. These features are unlikely to include information that will help differentiate the data. Another simple feature selection step is to remove features which are highly correlated. One of a pair of features that passes a set correlation threshold can be removed, as the information conferred is likely to be very similar. However, as correlation is linear, this method is not suitable for use if there may be useful nonlinear relationships within in the data.

Univariate feature selection is another commonly used technique that considers each feature separately. If the association between the outcome variable to be predicted and the feature is significant (p-value less than the specified threshold for significance), then the feature is included. The disadvantage of this method is a feature may become significant as it relates to other features, so there is the potential to lose important prediction information [149]. While multivariate approaches are more complex and computationally expensive than univariate approaches, it generates a feature set that arguably contains more informative patterns for the subsequent machine learning modelling. Forward and backward feature selection are two such methods. Forward selection begins with the most informative feature, and iteratively adds the feature with the next most information, up to the set number of features. Conversely, backward feature selection begins with all features, and iteratively removes the feature with the smallest amount of information in relation to all other present features. Both methods are referred to as “greedy” methods, because they evaluate one feature at a time [150].

LASSO, or Least Absolute Shrinkage and Selection Operator, regularisation is another method of feature selection that also regularises data at the same time [148]. The general idea of a regularisation method is to shrink the weight associated with features towards zero if they do not provide significant information for modelling. An upper bound is set for the total absolute value of the model features. This then requires that some features are regularised (shrunk) to meet this requirement. Unlike L2 regularisation, where only features shrinking can occur, LASSO regularisation will shrink features to zero if the cost of including them is too great, therefore excluding them from the future machine learning model. Methods based on machine learning algorithms such as linear support vector machines [151] and random forests [152] can also be used as feature selection methods. As well as the aforementioned supervised learning methods, unsupervised learning can be used to reduce the number of features. Methods such as principle component analysis attempt to condense a highly-dimensional set of data into a new set of fewer components. Rather than reducing features, this method creates a new set of features in lower dimensions that still accurately represents the structure and variation in the original data.

1.3.4 Overfitting

Models generated during supervised machine learning cannot be applied successfully in any circumstance unless they generalise well to new data. A model that is highly tuned to a specific data, for example because of the inclusion of too many features, will perform poorly due to overfitting. In the opposite case, if there is too little information to form a model that can sufficiently describe the relationships in the data, this model will not perform well on the training or test data, and is called underfitting. This dichotomy of under and overfitting is also known as the bias-variance trade-off. A high-bias model will have a high training and testing error due to a large gap between what is predicted and the truth. A high-variance model will have a high testing error because the model varied greatly in order to closely match the data points in the training data. In machine learning, the aim is to use the simplest possible model that performs to the standard expected. For example, choosing to use a fourth order polynomial equation that will give an accuracy of 0.80 on the training data, instead of a simpler cubic equation that gives an accuracy of 0.78, is a poor bias-variance trade-off that will likely result in a lesser performance when the model is applied to testing data.

1.3.5 Cross validation

As discussed in the previous section, training and testing a machine learning model can reveal whether a model has been overfitted to a dataset. A model needs to generalise well to other data in order to perform correct predictions. Cross validation is a technique used to subsample data in supervised machine learning, and summarised in Figure 12. This process is an extension of the basic two-fold split of the full dataset into the training and testing datasets, sometimes known as hold-out validation. The dataset is first partitioned into training and testing data, with the testing data being set aside and not used until the model is fully trained. This is crucial in order to objectively view how generalisable the model is to new data. Then the training data is partitioned into N folds of data. Popular splits are 3, 5 and 10-fold cross validation, or a leave-one-out approach, where the number of folds is equal to the number of samples in the training data. For each iterative round of model training, all but one fold is used to train the model, with the remaining fold used to validate the model. Each fold is used as validation data only once, and the overall performance of the machine learning model on the training data is the average performance over all validation folds. Then the trained model is applied to the testing data. Cross validation allows a user to train and validate a model several times over, while leaving the testing data completely unseen to the model. This allows an extra opportunity to check the generalisability of a model, without increasing the amount of data. Cross validation is more computationally intensive than simply splitting the data into training and testing sets. In the training of random forests, a similar process to cross validation called out-of-bag sampling is conducted, where a subset of the data is used to build different trees that subsequently assemble into a random forest model. For this reason, cross validation is not used when building random forest-based models.

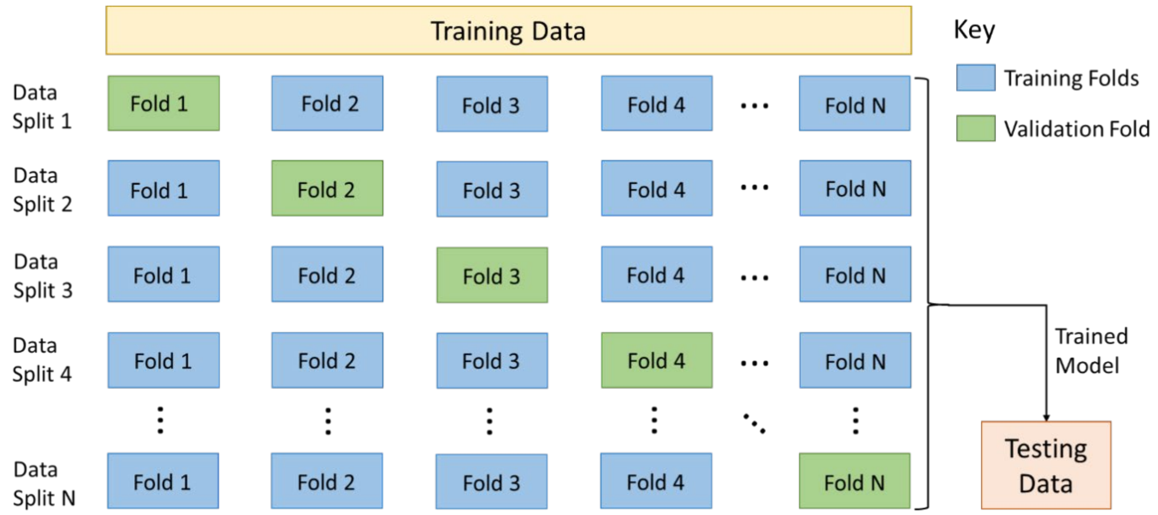


Figure 12 Schematic of the implementation of cross validation in machine learning.

1.3.6 Evaluation metrics

When evaluating the performance of a machine learning model, there are several metrics that can be employed. Two of the key metrics that give an overview of model performance are the area under the receiver-operator curve (AUC), and the F-score. The F-score, sometimes called the F_1-score or F-measure, is a weighted average of two metrics, precision and recall (sensitivity). AUC combines sensitivity and specificity. The AUC metric is unaffected by imbalanced data, which is common in machine learning. The F-score can be affected by skews in the number of data points associated with each outcome variable. However, there is evidence that the same AUC can produce different precision-recall curves, which supports a need to look at several evaluation metrics when evaluating model performance [153]. Many of the popular evaluation metrics are defined in Table 2.

Table 2 Definitions of commonly used evaluation metrics in machine learning.

Metric	**Definition**
Accuracy	Usually given as a percentage, a measurement of the total number of correct predictions made by a model.
Area under the receiver-operator curve (AUC)	Reported as a percentage or number between 0 or 1, this metric is calculated using the model sensitivity/recall and specificity
Balanced Accuracy	Usually a percentage, and should be used for imbalanced datasets. The total number of correct predictions in each class is weighted according to the proportion of data available in each class.
F-Score	A model performance measure (between 0 and 1 or a percentage) calculated using the recall and precision metrics.
Out-of-bag Error	A metric (between 0 and 1) exclusive to random forest-based methodologies, measuring the test error of an assembled model.
Precision	A measure also known as the positive predictive value, the number of true positives as a fraction of the total given a “positive” label (given either as a percentage or a number between 0 and 1).
Recall	This metric is also known as sensitivity.
R^2	Measurement of the variation in the data explained by a regression model.
Sensitivity	Given either as a percentage or a number between 0 and 1, this measures the number of correctly identified true positives.
Specificity	Given either as a percentage or a number between 0 and 1, this measures the number of correctly identified true negatives.

1.4 Thesis outline and aims

In this thesis I present work detailing the application of bioinformatic, statistical and computational methods with the aim of classifying or stratifying IBD patients, in order to further progress towards the ultimate aim of personalised medicine for individuals with this chronic, complex disease. This thesis first assesses the state of the field through a systematic review of ML applications to autoimmune disease. This topic is later revisited to gauge how this area of research as change, specifically for IBD. Using oxidative stress and antioxidant potential assay data, the connections between these markers, clinical data and genomic variation are elucidated. The main focus of the thesis is to develop optimal strategies to utilise genomic data alongside ML to classify IBD patients by their disease subtype, and CD patients by the presence or absence of a stricturing endotype. This included deducing the best way to prepare WES data to be used as input in ML algorithms and improving on feature selection processes. This was with the overarching aim of using ML techniques to bridge the gap between generating genomic data on patients with life-long clinical needs, and enabling this data to inform clinical management of these patients.

Chapter 2 Systematic review of the applications of artificial intelligence and machine learning for autoimmune disease

Chapter summary – the systematic review in this chapter uses a straightforward search strategy to assess artificial intelligence and machine learning (ML) applications to some of the most common autoimmune diseases (search performed December 2018). The aim was to evaluate the popular research questions for ML, which algorithms were most frequently used, and what types of data were common. A summary of these questions for each autoimmune disease is provided, alongside statistics such as the median sample size. The remainder of the results section is organised according to the research question (for example diagnosis, or autoimmune disease management). This work gave a broad overview of the current research in this interdisciplinary field.

Chapter contributions – systematic search performed by Imogen Stafford. Imogen Stafford and Melina Kellermann were first and second reviewers, respectively, for the assessment of study abstracts. Enrico Mossotto also assisted with study inclusion and exclusion in cases where this decision was challenging. Imogen Stafford gathered data from all papers and performed all further analysis.

2.1 Introduction

2.1.1 Autoimmune disease

Autoimmune diseases are chronic and complex, whereby genetics, the environment, and immune system dysregulation all contribute to their development (Figure 13). Due to the heterogeneity of onset and progression, diagnosis and prognosis for autoimmune diseases is unpredictable. The prevalence of autoimmune disease is difficult to estimate as diseases are variably represented across studies and no definitive list exists [154-156]. The approximate prevalence is evaluated to be between 4.5% [155] and 9.4% [154].

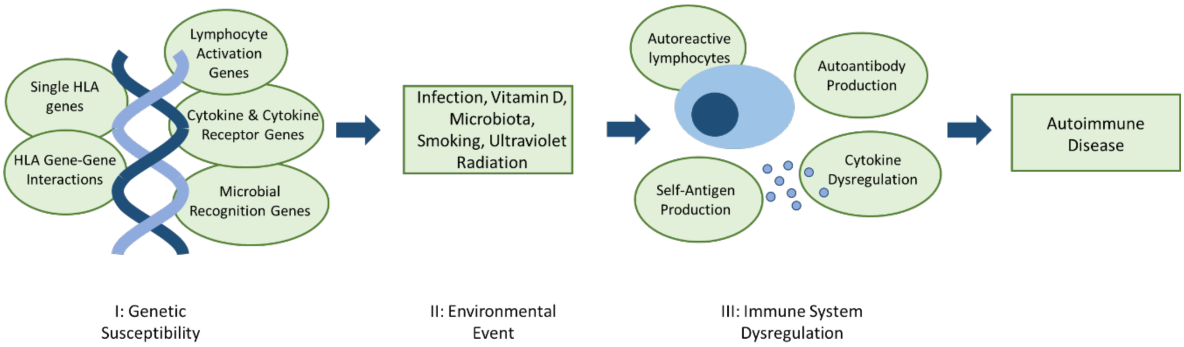

Lymphocyte Activation Genes
Single HLA genes
Cytokine & Cytokine Receptor Genes
HLA Gene-Gene Interactions
Microbial Recognition Genes
Infection, Vitamin D, Microbiota, Smoking, Ultraviolet Radiation
Autoreactive lymphocytes
Autoantibody Production
Self-Antigen Production
Cytokine Dysregulation
Autoimmune Disease
I: Genetic Susceptibility
II: Environmental Event
III: Immune System Dysregulation

Figure 13 Three factors that contribute to autoimmune disease development. I: Genetic predisposition is often conferred by a combination of genes that may include human leukocyte antigen (HLA) genes in the major histocompatibility complex (MHC). These directly or indirectly affect immune system regulation. II: examples of potential environmental events that trigger or contribute to dysregulation of the immune system. III: autoantibody production by itself will not always result in development of autoimmune disease, other dysregulation such as self-antigen production and unnecessary escalation of immune response mechanisms is often required [157]

The contribution of genetics towards autoimmune disease development has been illustrated with monozygotic and dizygotic twin studies. For example, the concordance of multiple sclerosis was estimated to be 25-31% in monozygotic twins, and 3-5% in dizygotic twins [158]. However, the range of autoimmune disease concordance in monozygotic twins was wide, 12-15% for rheumatoid arthritis in comparison to 75-83% for coeliac disease [158], indicating that the extent that genetics contributes varies considerably. Additionally, HLA-DQ genetic markers have been associated with multiple autoimmune diseases [159]. This paints a complex picture of genetic involvement in autoimmune disease, without considering other contributory factors.

Genetics often contributes to autoimmune disease by predisposing individuals to autoimmunity [160]. Resultant specific autoantibodies from loss of self-tolerance have been detected in patients before clinical onset in many autoimmune diseases [161]. Autoimmune disease will only manifest after further dysregulation in both the innate and adaptive immune system [162]. Microbial antigens, foreign antigens and cytokine dysregulation, can cause induction of self-reactive lymphocytes [157]. Hyper-activation of T and B cells may occur, along with a change in the duration and quality of their response which further disrupts the homeostasis of the immune system [162].

Gene-environment interactions can also contribute to autoimmune disease, through epigenetic mechanisms. Many environmental factors, including infections, ultraviolet light, environmental pollutants, smoking and diet, can induce epigenetic changes [163]. This can modify gene expression and also contribute to loss of tolerance [164]. Specific DNA methylation and histone modifications have already been identified for a number of more prevalent autoimmune diseases [165].

2.1.2 Personalised medicine

Personalised medicine is an area generating increasing interest, given its success transforming cancer treatment for some patients [166]. Application of these kinds of strategies are becoming achievable given current technologies. These approaches may be of particular value for complex diseases, such as autoimmune diseases. There is distinct variability within disorders [167], and a proportion of patients have additional autoimmune diseases (Table 3) due to shared developmental mechanisms [168]. Arguably, a 'one-size-fits-all' approach to treatment is not appropriate for this heterogeneity within diseases coupled with autoimmune co-morbidities. The realisation of personalised healthcare would lead to treatment of the causal molecular mechanism, resulting in better patient outcomes.

Table 3 Number of patients with one or more additional autoimmune diseases. These studies perform their analysis by first identifying a cohort with one autoimmune disease (left column), and subsequently reviewing the presence of other autoimmune diseases in the cohort.

Autoimmune disease	**Patients with additional autoimmune disease(s) (%)**
Rheumatoid Arthritis	24.3 [168]
Myasthenia gravis	15 [169]
Hashimoto's Thyroiditis	29.4 [170]
Vitiligo	19.8 [171]

Standard patient care generates a diversity of clinical data types, and these data are often accumulated longitudinally over the disease course. Examples include: images obtain during colonoscopies and magnetic resonance imaging (MRI), laboratory test results from blood or urinary samples, symptoms at diagnosis, successful and unsuccessful treatments, and time between flare-ups of a relapsing-remitting disease. Along with demographic data, this information is increasingly stored in electronic medical records (EMRs) [172], establishing these records as a rich data source.

In addition to a wealth of clinical data, 'omic data is becoming widely available. 'Omic data sets are large, as molecular measurements are made on a genome-wide scale [173]. It is sizeable enough that computational power and capacity remain a limitation [174], along with the expense of data storage and operation of these technologies, despite their high-throughput nature [175]. The advent of high throughput technologies has allowed quick analysis of many ‘omic data types, including the genome, transcriptome and proteome. Layering multiple sets of ‘omic data may give a fuller picture of the molecular status of individual’s autoimmune disease, leading to novel insights that could evolve into treatment strategies.

This wide variety of data types has limited clinical utility without methods for interpretation. There is a clear need for automated, intelligent systems, and computational tools that can uncover obscure, clinically relevant patterns within the wealth of data. Artificial intelligence and machine learning methods have the capacity to fulfil this purpose [176]. The ability to stratify patient’s using these data has implications for their care, from estimation of autoimmune disease risk, diagnosis and prognosis to management, monitoring and treatment response.

This systematic review aims to appraise the current applications of artificial intelligence and machine learning methods to autoimmune disease for improved patient care. The study identifies the most common models, data and application types. Potential areas for improvement in this area of exciting interdisciplinary research are established, and promising future possibilities discussed.

2.2 Methods

2.2.1 Autoimmune disease selection

The autoimmune diseases selected for the systematic review were based on prevalence estimates [154], choosing those that were most likely to have sufficient data for analysis using machine learning. These included: Addison disease, alopecia, Coeliac disease, Crohn’s disease, ulcerative colitis, type 1 diabetes, autoimmune liver diseases, hyper- and hypo-thyroidism, multiple sclerosis, myasthenia gravis, polymyalgia rheumatica, psoriasis, psoriatic arthritis, rheumatoid arthritis, Sjögren syndrome, systemic sclerosis, systemic lupus erythematosus, systemic vasculitis, uveitis and vitiligo.

2.2.2 Systematic literature search

Literature searches were performed electronically with OvidSP on the MEDLINE from 1946, and EMBASE from 1974 databases. An additional search on the Computers & Applied Sciences Complete database, available on EBSCO, was performed. This was to ensure the capture of all relevant studies, those aiming to solve medical problems, and those focusing on algorithm development that may use medical data. The literature search was completed in December 2018, last search 17/12/2018. Autoimmune diseases were searched for separately, using a search structure of the worlds “machine learning” or “artificial intelligence” combined with the selected search terms for that autoimmune disease. The diseases and their corresponding search terms are listed in Table 4. Boolean operators OR and AND were used to combine search terms systematically. In both databases, the search terms needed to be present in the title, abstract, or subject terms/keyword headings assigned by the study’s authors.

Table 4 Search terms used in OvidSP and EBSCO for each autoimmune disease.

Autoimmune Disease	**Disease Search Term(s) Used**
Addison’s Disease	Addison*
Alopecia	Alopecia
Celiac Disease	Celiac, Coeliac
Inflammatory Bowel Disease	Inflammatory Bowel Disease, Crohn* Disease, Ulcerative Colitis
Type 1 Diabetes	Type 1 Diabetes, Insulin?dependent Diabetes
Autoimmune Hepatitis	Autoimmune Hepatitis, Chronic Active Hepatitis, Primary Biliary Cirrhosis, Primary Sclerosing Cholangitis
Thyroid Disease	Autoimmune thyroiditis, Hashimoto* Thyroiditis, Hashimoto* Disease, Grave* Disease, Hyperthyroid*, Hypothyroid*
Multiple Sclerosis	Multiple Sclerosis
Myasthenia Gravis	Myasthenia Gravis
Polymyalgia rheumatica	Polymyalgia rheumatica
Psoriasis	Psoriasis
Psoriatic arthritis	Psoriatic arthritis
Rheumatoid Arthritis	Rheumatoid Arthritis

Sjögren syndrome	Sjogren syndrome
Systemic sclerosis	Systemic sclerosis
Systemic Lupus Erythematosus	Lupus
Systemic Vasculitis	Polyarteritis nodosa, microscopic polyangiitis, granulomatosis with polyangiitis, eosinophilic granulomatosis with polyangiitis.
Uveitis (iridocyclitis)	Uvetitis, iridocyclitis
Vitiligo	Vitiligo

2.2.3 Inclusion and exclusion criteria

Studies that applied machine learning methods to any autoimmune disease listed above, or to complications arising from autoimmune disease were included. Studies that applied machine learning to a non-autoimmune disease comorbidity in patients with autoimmune disease were excluded. Other applied exclusion criteria were: studies not written in English, a publication date before 2001, machine learning not trained on real, human patient data, articles that were not peer-reviewed, and review articles. This systematic review conforms to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standards [177].

2.2.4 Data Visualisation

Studies were assigned an ML type according to the ML method used. Studies were counted multiple times if there was more than one ML method recorded. The ML type and the corresponding autoimmune disease this was applied to, and the IBD clinical applications ML was applied to, was plotted in R [178] using ggplot2 [179].

2.3 Results

2.3.1 Summary of results

A total of 702 papers were identified in database searches, of which 169 met the criteria for inclusion in analysis. 227 duplicate records were removed, 273 records were excluded after reading the abstract, and 33 excluded after a full text read (Figure 14). Information on the included studies is summarised in Table 5, and a more detailed breakdown of the contents of each study is given in Supplementary Table 2. Of the autoimmune diseases included in searches, six did not return any study that met criteria for analysis: Addison disease, myasthenia gravis, polymyalgia rheumatica, Sjögren syndrome, systemic vasculitis, and uveitis.

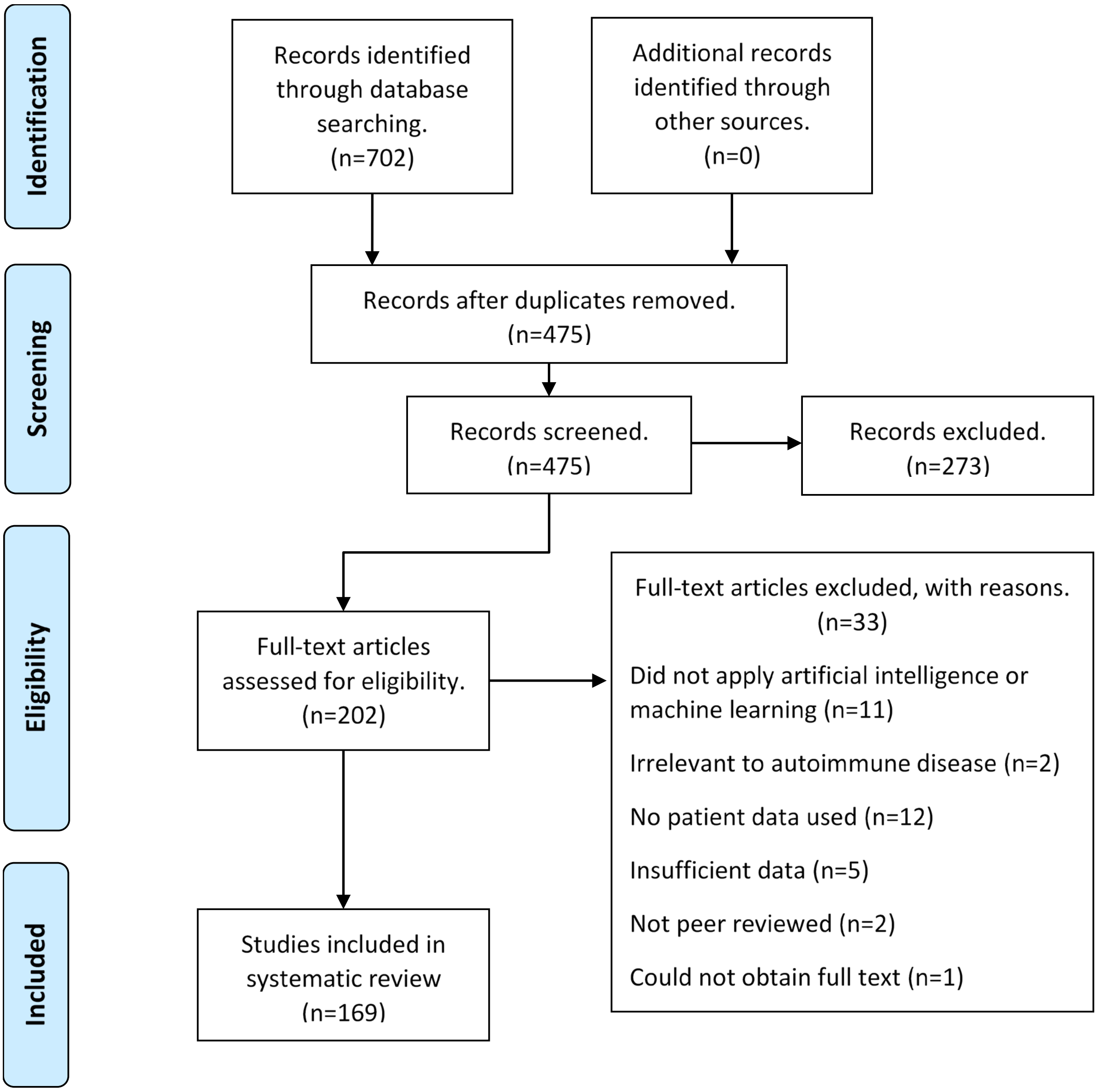

Identification
Screening
Eligibility
Included
Records identified through database searching. (n=702)
Additional records identified through other sources. (n=0)
Records after duplicates removed. (n=475)
Records screened. (n=475)
Records excluded. (n=273)
Full-text articles assessed for eligibility. (n=202)
Full-text articles excluded, with reasons. (n=33)
Did not apply artificial intelligence or machine learning (n=11)
Irrelevant to autoimmune disease (n=2)
No patient data used (n=12)
Insufficient data (n=5)
Not peer reviewed (n=2)
Could not obtain full text (n=1)
Studies included in systematic review (n=169)

Figure 14 Flowchart recording number of papers reviewed in each stage. During the screening and eligibility stages the inclusion and exclusion criteria are applied, first to the title and abstract, and subsequently to the full text. For some records at the screening step, inclusion or exclusion could not be established based on abstract only, and so a full read of the record was completed at the eligibility stage. Two reviewers screened records independently. If consensus on the record could not be established, a third reviewer assessed the article and determine whether it was included or excluded.

Table 5 Artificial intelligence and machine learning applications to autoimmune diseases. Study information is recorded per autoimmune diseases and includes: study count, year range studies were published in, popular applications and methods, and all data types used. Median (range) sample size is included rather than the mean, due to the inclusion of a minority of studies that had very large cohorts, usually analysing genome wide association study data, or electronic medical records.

Disease	**Number of Studies**	**Years**	**Most Popular Classification/Prediction Application(s)**	**Most Popular Machine Learning Method(s)**	**Median Sample Size (min, max)**	**Data Types Used**
Multiple Sclerosis	41 [180-220]	2008-2019	Diagnosis, Prognosis, Disease Subtype	Type of Regression, Random Forest, Support Vector Machine	99 (12, 12566)	Clinical, Survey, Genetic, MRI, Lipid Markers, SNPs, Gait Data, Immune repertoire, Gene Expression
Rheumatoid Arthritis	32 [221-252]	2003-2018	Risk, Diagnosis, Early Diagnosis, Identify Patients	Support Vector Machine, Variations of Random Forest, Neural Network and Decision Tree	338 (22, 922199)	Medical Database, Immunoassay, Metagenomic, Microbiome, GWAS/SNP, Clinical, Movement Data, Amino acid analytes, Transcriptomic, EMRs, Ultrasound images, Proteomic, Laser images
Inflammatory Bowel Disease	30 [253-282]	2007-2018	Diagnosis,	Random Forest, Support Vector Machine	273 (50, 53279)	Clinical, Colonoscopy Images, Metagenomic, Gene Expression,

			Response to Treatment, Disease Risk, Disease Severity			GWAS, Microbiota, miRNA Expression, EMRs, Exome, MRI
Type 1 Diabetes	17 [283-299]	2009-2018	Disease Management	Novel Methods/Hybrid Models, Neural Network, Support Vector Regression	23 (10, 10579)	Clinical, Red Blood Cell Images, VOCs, GWAS/SNPs
Systemic Lupus Erythematosus	14 [300-313]	2009-2018	Variations of prognosis, Diagnosis	Logistic Regression, Neural Network, Random Forest Decision Tree	318 (14, 17057)	Clinical, Electronic Health Records, Drug Treatment, SNPs, MRI, Exome, Gene Expression, Proteomic, Urine Biomarkers
Psoriasis	11 [314-324]	2007-2018	Diagnosis, Disease Severity	Support Vector Machine	540 (80, 22181)	Digital Image, GWAS, Proteomic, RNA Biomarkers
Coeliac Disease	7 [325-331]	2011-2018	Diagnosis	Random Forest, Logistic Regression, Bayesian Classifier, Support Vector Machine, Logistic Model, Natural Language Processing, Combined Fuzzy Cognitive Map and	465 (47, 1498)	VOCs, Clinical, Peptide, EMRs

				Possibilistic Fuzzy c-means clustering.		
Thyroid Diseases	6 [332-337]	2008-2018	Diagnosis	Hybrid Models	215 (215, 7200)	Clinical
Autoimmune Liver Diseases	5 [338-342]	2009-2018	Prognosis	Variations on Random Forest	288 (64, 787)	Clinical, Clinical Trial, Microbiome
Systemic Sclerosis	4 [343-346]	2016-2018	Diagnosis, Treatment, Prognosis	Support Vector Machine, Random Forest	119 (37, 991)	Gene Expression, Nailfold capillaroscopy images, Peripheral Blood Mononuclear cell data (flow cytometry, DNA, mRNA)
Alopecia	1 [347]	2013	Comorbidity Analysis	Natural Language Processing	3568	Patient Data Repository
Vitiligo	1 [348]	2013	Comorbidity Analysis	Natural Language Processing	3280	Patient Data Repository

The diseases where machine learning and artificial intelligence techniques were most prevalent were multiple sclerosis (MS), rheumatoid arthritis (RA) and inflammatory bowel disease (IBD). These models used the highest variety of data. In addition, only models of these diseases used two data types (13/169 studies, clinical data was always one data type). Support vector machines and random forests were the most common machine learning methods, through all types of application and autoimmune diseases (Figure 15). The highest variety of machine learning types were applied to RA, followed closely by IBD. For every autoimmune disease clinical data was used in creating models, and for the majority of diseases a type of genetic data. The heterogeneity of machine learning methods and the pipelines they reside in, applications and data, as well as validation and evaluation of different approaches (Supplementary Table 2) renders a meta-analysis inappropriate.

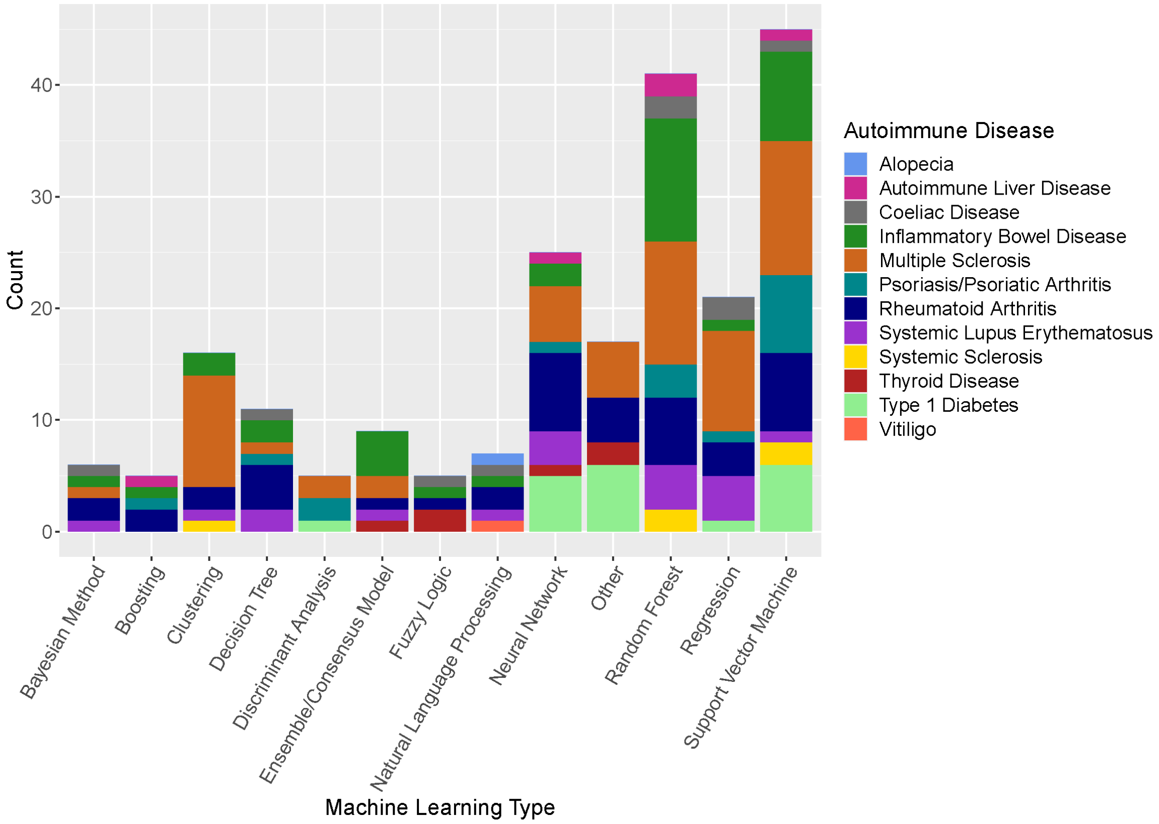


Figure 15 Stacked bar chart of types of machine learning found in the systematic review, grouped according to the main autoimmune disease they were used for. The other category includes uncommon methods, and novel ML pipelines.

The applications of machine learning to autoimmune disease can be categorised into six broad areas: risk prediction, patient identification, diagnosis, classifying disease subtypes, progression and outcome, and monitoring and management.

2.3.2 Identifying and assessing autoimmune disease risk

The two main applications in this category were disease risk prediction [211, 221, 238, 266, 275, 276, 282, 287, 291, 292] and identification of novel risk factors using feature selection [234, 240, 248, 259, 313] for IBD, type 1 diabetes (T1D), RA, systemic lupus erythematosus (SLE) and MS. Random forest, support vector machine and logistic regression were popular machine learning model types. A form of genetic data was utilised in fifteen studies, using sequencing arrays (GWAS), exome data (9 studies), gene expression data [211, 259], individual SNPs [291] within the HLA regions [191, 287], or from pre-selected genes [240]. Two studies combined a type of genetic data with clinical data [191, 211], and one used clinical data only [221].

2.3.3 Patient identification

The focus of this group of studies was to identify patients with autoimmune diseases from their electronic medical records using natural language processing [230-232, 301, 309, 329, 330]. Gronsbell et al. focused on increasing the efficiency of these types of models [245, 250]. The intention was for algorithms such as these tor replace International Classification of Diseases billing codes, which have reported error rates of 17.1-76.9% because of the inconsistent terminology used [301]. These algorithms also identify a cohort for further analysis, whether that be with machine learning or other methods. Natural language processing was also used in the two identified comorbidity studies for alopecia and vitiligo. Both diseases had similar autoimmune comorbidities [347, 348].

2.3.4 Diagnosis

Patient diagnosis was the most frequent application of machine learning, and used for all diseases. Support vector machines and random forests were the most frequently utilised model types for this area. Twenty-seven studies focussed on classifying cases and controls. This model type could have applicability in specific cases, for example where patients are asymptomatic. However, distinguishing cases from controls may not be as clinically useful as the other models in these studies, such as those using patients with a different autoimmune disease as controls [241-243, 308], exploring the classification of multiple autoimmune diseases [272, 341], or distinguishing diseases with similar presentations [209, 214, 236, 320, 326, 346, 349], for example coeliac disease and irritable bowel syndrome. Early diagnosis was specified as important for the degenerative conditions MS and RA, and so seven studies developed models for that aim [190, 192, 227, 243, 244, 249, 252]. More diagnostic applications included stratifying those with coeliac disease from an at risk group [325, 327], and distinguishing those likely to develop T1D complications [285, 295].

2.3.5 Classifying disease subtypes

Machine learning classified RA (one study), IBD (two studies) and MS (six studies) disease subtypes. These methods differentiated between CD and UC in the case of IBD, and two or more of the four MS subtypes: relapsing remitting MS, primary progressive MS, secondary progressive MS and progressive relapsing MS. Despite unsupervised methods being utilised infrequently, this area featured the use of three different unsupervised clustering algorithms: hierarchical clustering for identifying novel IBD subtypes [254]; consensus clustering to identify high, low and mixed levels of inflammation in RA [226]; and agglomerative hierarchical clustering to cluster MS by genetic signature [188]. Two of the previous studies also employed the supervised method support vector machines [226, 254]. There was a wide variety of data types considering the small number of studies: clinical (particularly MRI), genetic, RNA sequencing and gene expression data were all utilised.

2.3.6 Disease progression and outcome

Aside from diagnosis, predicting aspects of prognosis was the most prevalent area for model development. Twenty-seven studies focused on disease progression and patient outcomes. Other study emphases were disease severity [233, 265, 315, 316, 318, 321, 331] in psoriasis, RA, IBD and coeliac disease; treatment response [228, 239, 251, 258, 260, 261, 268, 273, 338] in IBD, RA and primary biliary cirrhosis (PBC); and survival prediction [247, 306, 342] in PBC, RA and SLE. Other models focused on improved image segmentation to aid prognoses [207, 213, 216, 280, 281, 312] for IBD and MS. Commonly used methods were support vector machines, random forests and neural networks. Few studies utilised ‘omic data [224, 256, 273, 324], with the majority using clinical data as a machine learning input.

2.3.7 Monitoring and management

Machine learning was used for the monitoring and management of T1D, MS and RA. Of the ten studies in T1D, four were for blood glucose predicted, four focused on predicting or identifying hypoglycaemic events and two used machine learning to support decision making using decision support systems or case-based reasoning. The majority of these models used clinical data. The other models were developed for monitoring movement in MS (three studies) and RA (one study) using activity measurements. Support vector regression was the most frequently used method [189, 199, 289, 290, 297].

2.3.8 Inflammatory Bowel Disease

As the research conducted in subsequent chapters centres IBD specifically, it was thought appropriate to summarise the systematic review findings in relation to this disease. While 11 different ML types were used in modelling for IBD, a large proportion of studies employed Random Forests (55%) [253, 256, 258, 261, 262, 268, 272-274, 280, 281] and Support Vector Machines (40%) [254, 255, 262, 264, 265, 267, 276, 278]. The clinical task types that ML was applied to for IBD are visualised in Figure 16. Another factor of interest was the composition of the cohorts used for these studies. Some studies had cohorts of CD [256, 263, 265, 271, 273, 275, 279-282]and UC [255, 257, 260, 261, 270] patients (33.3% and 16.7%, respectively), and others treated IBD as a singular disease group (20%) [253, 258, 262, 267, 268, 276]. The remaining studies made note of the CD and UC subgroups within their cohort for analysis. Finally, it was noted that there were 5 studies that utilised genetic data [266, 275, 276, 279, 282], and this data type usually comprised genetic array-based data or the inclusion of selected SNPs. Two studies had WES data available [275, 282], of which one used this data to impute genotypes [282].

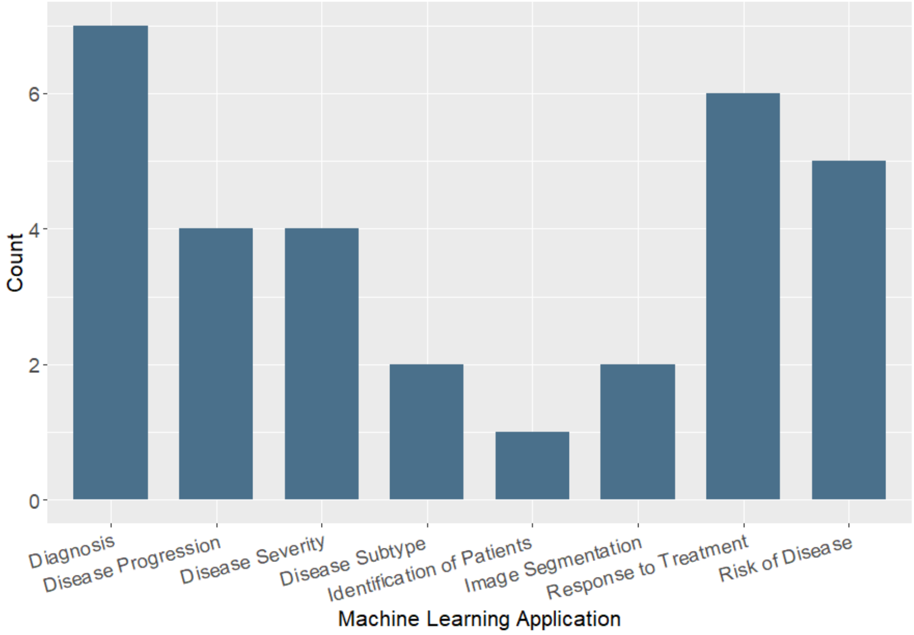


Figure 16 Number of studies per each prediction or classification task that ML was applied to for IBD.

2.3.9 Validation and independent testing

Of the 169 studies evaluated, 11 did not use any cross-validation method, so in these cases model robustness and applicability is uncertain. Not including research that used random forest models (where it is unnecessary to use cross-validation), or neural networks (where a cross-validation process can be too computationally intensive), 18/169 models only used hold-out validation. These models may be of clinical use, but unless the dataset is very large these methods have not been as robustly validated in comparison to those that use k-fold cross validation, a leave-one-out approach, or the combination of cross-validation methods and application of the method to an independent data set. A minority of studies (14/169) did use the latter combination for evaluating their models. This research did not have any machine learning algorithm types or applications in common, and the studies were for many different autoimmune diseases. The most common input data was clinical and genetic data.

2.4 Discussion

The variety of the machine learning models used and the pipelines that contain them reflects the heterogeneity of the autoimmune diseases the methods were utilised for. This makes it challenging to determine the methods that would be most effective, carried forward to further validation, and ultimately clinical application. Alternatively, instead of choosing one model, many could be combined with the aim of gaining consensus for the specific machine learning task. Modelling utilised an assortment of ‘omic data, including proteomic, metagenomics and genomic data. More common were sequencing array (SNP/GWAS) data, especially when the focus was predicting disease risk. Undoubtedly the most prevalent data type was clinical and laboratory data.

Data accessibility is critical for incorporating machine learning models into everyday clinical practice, and EMRs provide this for clinical and laboratory data. Some initiatives have moved to storing other data types in these systems, which will be essential for incorporating of multiple datatypes at a large scale. The eMERGE (electronic medical records and genomics) network integrates the genomic and EMR data repositories [350]. The SPOKE (Scalable Precision Medicine Oriented Knowledge Engine) study aims create an intelligent system that integrates data types in the storage platform, whilst analysing the connection between GWAS, gene ontology, pathways and drug data and EMRs using unsupervised machine learning [351]. Understanding the relationships between these and other data is key to implementing personalised medicine.

Personalised medicine approaches have already revolutionised cancer prognoses, improving patient outcomes and quality of life, accompanied by economic benefits to treatment providers.

Precision treatment has been propelled by the identification of cancer-specific driver mutations [352], allowing the identification of molecular diagnosis that subsequently influences the treatment strategy. Using targeted therapies, for example monoclonal antibodies and small molecule inhibitors has transformed the treatment of some cancers, or improved survival times [166]. Both cancer classification [353, 354] and pathway discovery has been achieved using machine learning. Classical treatment of autoimmune disease has usually involved a broad-brush approach to treatment. By utilising machine learning in conjunction with 'big' data, patients could be stratified into groups, and the appropriate treatment identified: the approach that has been effective in cancer. Currently, some studies have already exhibited this approach by using machine learning to investigate IBD subtypes [254], and stratifying inflammation status in an RA patient cohort [226].

Of the many models that were created for autoimmune disease diagnosis, usually classifying patients and controls, the majority achieved good classifier performance (where a combination of metrics are over the following thresholds: accuracy > 81%, AUC > 0.95, Sensitivity > 82, Specificity > 84). Although these classification tasks were somewhat simple, they illustrated machine learning's utility in diagnostics.

With respect to research specific to IBD, this systematic review identified some underexplored areas in the field. WES data was determined to be a rare data type to use [275, 282], and while there were 5 studies total utilising a form of genetic data, no study using this data type employed a Random Forest algorithm, which was very common in general for ML applications to IBD. As an algorithm that can leverage data containing non-linear relationships, Random Forest could allow the extraction of non-linear gene-gene interactions in genomic data for the benefit of IBD clinical classification tasks. Further, the majority of studies either considered IBD as a single disease class, or their cohort consisted of only CD, or only UC patients. This combined with only 2 studies building classifiers based on IBD subtype, suggests that a further look at using ML to analyse subtype differences could be beneficial to the field.

Six of 169 models from the literature returned more than one of the following metrics as either 1 or 100%: AUC, accuracy, precision and recall, sensitivity and specificity [186, 249, 285, 295, 317, 343]. A perfect performance indicates that a machine learning model may not be necessary, as the some variable(s) in the dataset classify the groups with no error. Alternatively, this performance may indicate overfitting without robust evaluation, or the poor implementation of cross-validation techniques.

When researchers reported machine learning results, the metrics used varied considerably, but often included accuracy, AUC, sensitivity and specificity. For the majority of machine learning tasks accuracy is an inferior measure to AUC, particularly when the dataset is imbalanced [355]. The AUC measure is not affected by an imbalanced dataset, but precision-recall curves may more accurately reflect the performance of a model [153]. In the case of creating and evaluating any model, it is important to decide the metrics that are most important to its evaluation. That is, whether to minimise the false positives or false negatives. Scully et al. illustrated this with their lesion segmentation model for SLE, which achieved a high specificity (99.9%) by labelling all tissue as non-lesion [312].

A small proportion of the studies combined cross validation with a separate testing set for more robust model evaluation, and the importance of this was demonstrated with Ahmed *et al's* [227] machine learning model. In their study the AUC dropped by 0.25 using an independent dataset, indicating a decreased model performance on new data, and the importance of an independent dataset to assess the generalisability of a model.

The literature reviewed here demonstrated that artificial intelligence and machine learning methods can provide useful insight, and potentially improve patient outcomes, despite the heterogeneity of autoimmune disease presentation, diagnosis, and prognosis. The diversity in data used, machine learning models, and in particular model evaluation, is a preventative barrier to transferring the knowledge obtained with these models to the clinical practice. Further, the focus of the systematic search was restricted to a chosen list of autoimmune diseases, which may have not fully captured all literature using machine learning for autoimmune diseases.

From consideration of the studies included here, it appears appropriate to advocate for the standardisation of model evaluation, a combination of cross validation and independent test data for model validation. Results should be reported using the full spectrum of evaluation metrics, including AUC, sensitivity, specificity and F1 score. Increased confidence in model results may allow for more complex model creation, through layering data types, or combining models. These methods could then by applied to tasks that mirror the complexity of autoimmune diseases. Through these improvements, artificial intelligence and machine learning brings the reality of personalised medicine closer for, not only patients with autoimmune disease, but those with any common, complex disease.

Chapter 3 Methods and method development

Chapter summary – this chapter discusses the methods used for the processing and transformation of the WES data used throughout this thesis. This includes the alignment, variant calling and annotation of individuals alongside the quality control of WES batches. In addition, the processing of this data as a cohort – variant joint calling, filtering for a high quality callset and subsequent annotation – is also outlined. The transformation of WES data into a matrix of per-gene, per-patient scores (called GenePy scoring) is detailed. Over the course of my PhD project, it was necessary to implement bioinformatic pipeline upgrades for the joint calling process, annotation and transformation of data into GenePy scores. Bioinformatic pipeline upgrades are a mainstay of genomic informatics, and this represented a substantial component of research time. For this reason, both the original (used in Chapter 4) and upgraded pipelines (used in Chapter 5 and onwards) are described. Methods that are specific to each chapter are discussed within their respective chapter.

Chapter contributions – initial pipelines for alignment, variant calling and annotation of WES data for individuals and the IBD cohort were run by Imogen Stafford. Quality control was performed by Imogen Stafford on three batches of WES data, with other batches quality controlled prior to this thesis, and quality control of the 2020 batch of adult IBD data performed by Guo Cheng. Guo Cheng and Imogen Stafford aligned and called individuals for the updated pipeline, Guo Cheng performed the joint calling of all individuals. Imogen Stafford developed, implemented and documented new annotation and GenePy scoring processes.

Supplementary files can be found at https://doi.org/10.5258/SOTON/D2655. Throughout the Chapter relevant files and scripts are referenced to ensure reproducibility of whole exome sequencing data processing. Static versions of GitHub repositories where joint calling pipelines and GenePy scoring pipelines are detailed in full are included in Supplementary files.

3.1 Introduction

Raw data generated by high-throughput sequencing need to undergo a series of processes to extract clinically significant information. Generating files ready for analysis consists of three stages (Figure 17): alignment of data to the reference genome; calling for sites where the sample data differs from the reference genome (variant calling); and annotation, where additional information regarding each variant, such as allele frequencies and predicted deleteriousness metrics, are added. Some of the files created during variant calling and annotation are used for quality control. Quality control is essential to 1) check that data of sufficient depth has been received; 2) ensure that identifiers associated with the sequencing data are correct and have not been swapped, and 3) to check for contamination, either from another sample, or outside sources. Probands in each batch are checked to ensure sufficient depth of coverage, the BAM (binary alignment map) file size is correct, and that the clinically recorded sex matches with the genetics. Checking the percentage of shared variants between individuals in the batch is used to cross-check the number of related individuals (this can indicate contamination from other DNA). A selection of 24 SNPs for each proband are also tested separately [356], and the genotypes of these sites cross-checked with the sequencing data. Sequencing data can be processed for each individual, or the data can be analysed as a cohort, through creation of a multi-call variant call format (VCF) file at stage two, followed by annotation.

The traditional annotated VCF file created from sequencing data is not ideal as an input for machine learning. The number of variants per patient would lead to a highly-dimensional dataset, increasing in size as very rare and private variants are added with each patient sample. Furthermore, the standard VCF file does not benefit from additional information such as variant deleteriousness metrics and allele frequencies that can add biological and clinical meaning to sequencing output. For this reason, a key method for further analysis in this thesis was the generation of a GenePy matrix. GenePy [357] is a tool that creates a per gene, per individual score based on the number of variants a patient has per gene, integrating the zygosity, minor allele frequencies and predicted deleteriousness of those variants. This gene-level scoring approach is particularly valuable for complex diseases such as IBD. The causes of such diseases may be compound heterozygous variants, or the additive effect of many variants across one or more genes. The GenePy matrix forms a standardised input suitable for integrating with other data sets and machine learning. The three main pipelines for data processing are summarised in Figure 17.

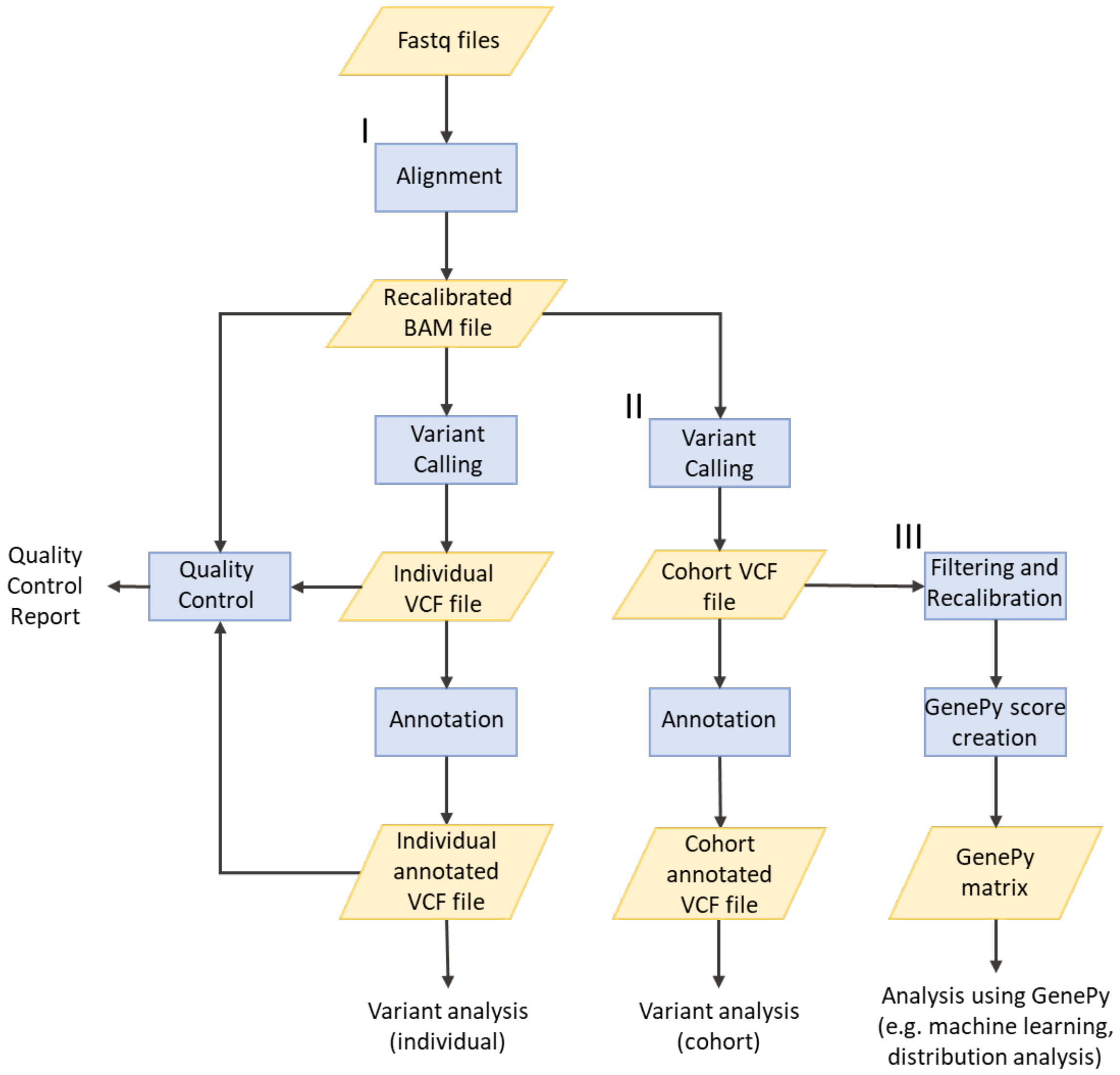

Fastq files
I
Alignment
Recalibrated BAM file
Variant Calling
II
Variant Calling
Quality Control Report
Quality Control
Individual VCF file
Cohort VCF file
III
Filtering and Recalibration
Annotation
Annotation
GenePy score creation
Individual annotated VCF file
Cohort annotated VCF file
GenePy matrix
Variant analysis (individual)
Variant analysis (cohort)
Analysis using GenePy (e.g. machine learning, distribution analysis)

Figure 17 Three main exome sequencing data processing pipelines. I) Processing of individual's exome data. This is typically done as batches are sequenced, and files created during the process are used to assess quality. The annotated VCF is used to assess potential causal variants on an individual basis. II) Variant calling and genotyping VCFs to create a cohort file (joint calling) that can be annotated to analyse variants on a cohort basis. III) Steps for the creation of a GenePy matrix, based on a cohort VCF. These scores can be used in different types of analysis, for example machine learning.

3.2 Programming and bioinformatic resources

3.2.1 Iridis 5

Iridis 5 is the latest generation of the University of Southampton's high performance computing cluster, and all exome sequencing data processing was completed with the use of this system. Iridis 5 is four times more powerful than the previous system: it has 464 computing nodes with 40

CPUs and 192GB of memory per node. Additionally, there are four high-memory nodes with 64 cores, 768GB of memory and 9TB of local temporary storage space. Over 20,000 processors provide 1,305 TFlops peak. Supercomputers like Iridis 5 are important to facilitate fast and efficient processing of increasingly large volumes of sequencing data. Processing data is executed using the bash command line which is based on the Unix architecture.

3.2.2 Burrows-Wheeler Aligner

The Burrows-Wheeler Alignment tool (BWA) implements the Burrows-Wheeler Transform (BWT) algorithm for the alignment of sequencing reads [80, 358]. BWT was originally developed for the compression of text string data, with a key factor being that additional data does not need to be stored to reverse compression. In genomics, BWA uses this compression for quick alignment, as each read is essentially a text string that often contains many repeats [80]. The BWA software has three different aligners: BWA-backtrack, BWA-MEM and BWA-SW. The former is designed for shorter reads (less than 100 base pairs), and latter two can align reads ranging from 70 base pairs to over a megabase. BWA-MEM is the faster and more accurate of the two, and so is the tool utilised in the bioinformatic pipelines in this chapter. In comparison to other aligners currently available, BWA-MEM is not as accurate as Novoalign, however it is much faster [358]. When the alignment of large genomic datasets is required, BWA-MEM represents the best trade-off between accuracy and speed.

3.2.3 Genome Analysis Toolkit

Developed by the Broad Institute, the Genome Analysis Toolkit (GATK) is a software package that can implement a number of tools for processing sequencing data [359]. GATK is used in the bioinformatic pipelines detailed below to complete two key processes: variant calling and variant quality score recalibration. Variant calling in diploid organisms is completed using HaplotypeCaller. This software is very popular because variant calling can be scaled up to include more samples without losing accuracy or sensitivity [360]. During variant calling, when HaplotypeCaller encounters a region with variation, it ignores the existing alignment and reassembles the reads in that region. This process means increased accuracy, particularly in regions with a lot of variation, and an increase in the calling of insertions and deletions [359]. By design, HaplotypeCaller is very sensitive in order to achieve the maximum number of variant calls. For users who want to filter the variants called for an overall higher quality call set with fewer false positives, GATK implements variant recalibration (VariantRecalibrator, ApplyRecalibration). The first step uses machine learning to assign a probability score of a variant being a true positive, and the second filters the call set according to the sensitivity required of the call set (i.e. the balance between missing true variants and including false positives [359].

3.2.4 ANNOVAR

The software package ANNOVAR is used to annotate variants to interpret their consequences [361]. ANNOVAR only requires the VCF file and text files of any supported database to annotate variants. The three main types of annotation are gene-based, filter-based and region-based annotation. Gene-based annotation provides information regarding a variant's position in the genome (e.g. exonic, intronic, close to a splicing site), and if the variant is exonic, what type of variant it is (e.g. nonsynonymous SNV, frameshift deletion). In filter-based annotation, the specific variant is searched for in the chosen annotation databases. This type of annotation can provide information about the frequency of a variant in a population, and its likely deleteriousness. Region-based annotation will not search for the specific variant but instead a region which can include one or more bases (e.g. chromosome 1:1000-1000, chromosome 3:2000-2050). The nucleotide change is not important in region-based annotation.

3.2.5 Ensembl Variant Effect Predictor

The Ensembl Variant Effect Predictor (Ensembl-VEP), is another annotation tool similar to ANNOVAR, which is available through the Ensembl website, or to download for offline use [362]. Input files can be in many formats, including a white space separated file of variants, and a VCF file format, and can be output as a tab-delimited, VCF, or JSON (JavaScript Object Notation) format. A file can be annotated based on Plugins, or a custom annotation. Plugins are supported by Ensembl-VEP, and they are downloaded, along with the related reference database, in order to annotate the file. Ensembl-VEP's custom annotation system means that any file can be used to annotate the input, including BED files and VCF files. Ensembl-VEP has a more granular annotation of variant consequences when compared to ANNOVAR. For example, where all splicing variants would be labelled as "splicing" by ANNOVAR, Ensembl-VEP categorises these into "splice donor" (splice variant at the 5' end), "splice acceptor" (splice variant at the 3' end) and "splice region" variants. Further, the transcripts used for annotation can be specified (for example the canonical transcript), and multiple annotations per variant (based on multiple transcripts) can be reported.

3.2.6 CADD

The Combined Annotation-Dependent Depletion (CADD) score is a measure of variant deleteriousness that is able to score single nucleotide variants and short insertions and deletions

[88]. CADD does not use any prior knowledge of variant pathogenicity in its scoring. Instead, it employs machine learning to predict pathogenicity. The early versions of CADD employed a support vector machine, but versions 1.4 and 1.5 use a logistic regression model. The model is trained on millions of variants split into two groups: real “proxy-neutral” variants that have become fixed in the genome and therefore are mostly benign and simulated “proxy-deleterious” variants that are *de novo* and free of selective pressure. A variant that appears closer to the simulated scenario is then presumed to be likely deleterious, as this variant would not become fixated. The advantage of this approach is that all variants can be scored (approximately 9 billion potential single nucleotide variants) [88]. The most recent version of CADD, version 1.6 or CADD-Splice [363], also incorporates information from the bioinformatic tools for scoring splicing variants, MMSplice [364] and SpliceAI [365].

3.2.7 Genome Aggregation Database

The Genome Aggregation Database (gnomAD) version 2 is a database containing 125,748 exomes and 15,708 genomes assembled by the Broad Institute, mapped to the GRCh37 reference sequence [82]. The sequences are derived from individuals with 6 global and 8 sub-continental ancestries. The database is an excellent source of information for the expected frequencies of variants in a population. After data processing, 14.9 million and 229.9 million high-quality variants were identified in the exome and genome datasets, respectively. The more recent version 3 consists of 76,156 whole genomes mapped to GRCh38, however, this database is not as powered for annotating coding variation as version 2, and as such it is still recommended to use the version 2 databases lifted over onto GRCh38 if analysis only involves the exonic regions.

3.2.8 GenePy

GenePy is a gene pathogenicity scoring system which assigns one score to each gene that reflects the pathogenicity conferred by all variants present in that gene [357]. Each GenePy score is calculated with the following equation

$$S_{gh} = \sum_{i=1}^{k} D_i \, log_{10}(f_{i1} \cdot f_{i2})$$

Where the score *S* is calculated for each gene *g*, and individual *h*. The frequency of both alleles at each locus *i* are represented by f_{i1} and f_{i2}. This is multiplied by the deleteriousness metric *D* for every locus. Although any allele frequency database and deleteriousness metric can be employed to construct GenePy scores, here the aforementioned gnomAD [82] and CADD [88, 363] were utilised. This enables the generation of a gene-by-individual matrix. There were two main reasons for using this scoring system. Firstly, an annotated VCF file of all variants present in a cohort would be highly-dimensional, and the consequent data sparsity means it is more difficult to produce informative ML models in downstream analysis. Processing the data so it is at the gene level, rather than the variant level therefore reduces the dimensionality of the dataset. Secondly, in initial studies GenePy score distributions were found to be significantly different in cases and controls for an IBD dataset, and a Parkinson's disease dataset (examples of GenePy score distributions for the IBD cohort utilised in this research are shown in Figure 18) [357]. These distribution differences between groups can be leveraged through ML. Of particular value in the calculation of GenePy scores is its ability to summarise rare and common variation in an individual, ideal for a cohort of IBD patients, as the genetic makeup of individual's disease may range from monogenic to polygenic, as described in Section 1.2.

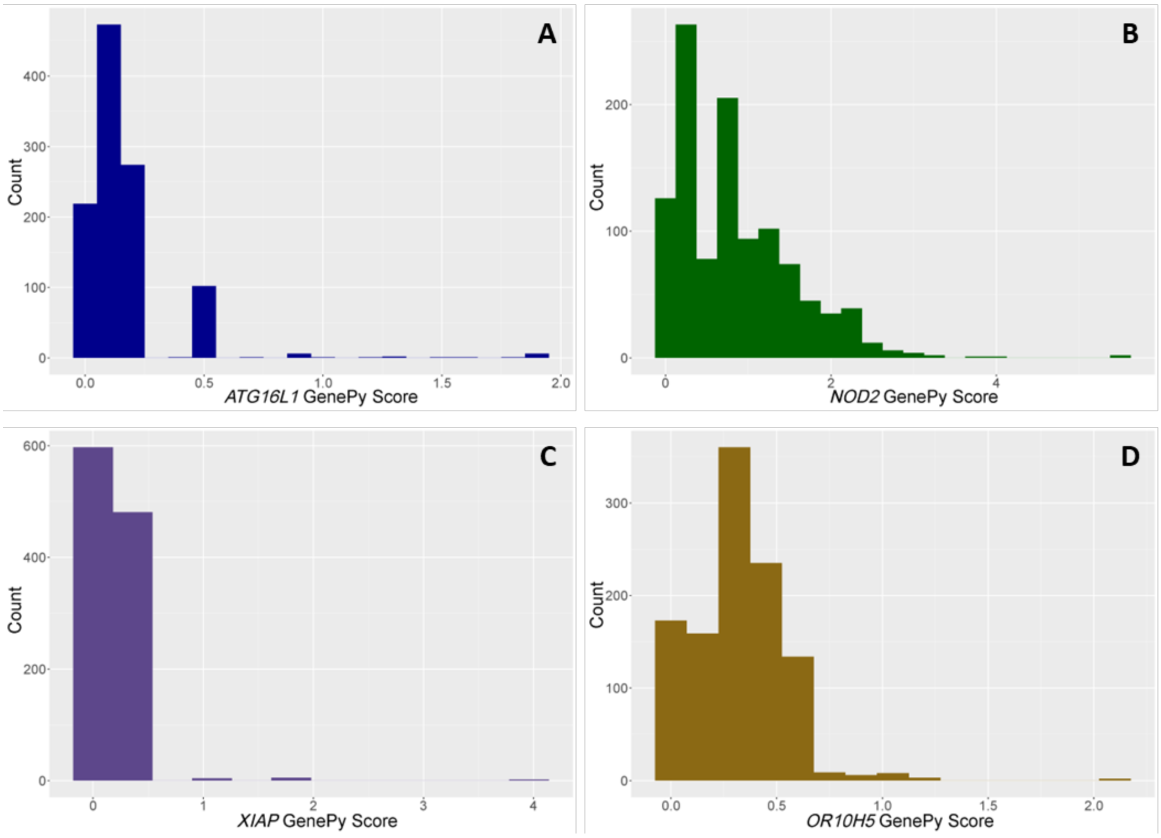


Figure 18 Examples of GenePy score distributions within the IBD cohort, which highlights varied distributions and ranges of GenePy scores. A) *ATG16L1*, a gene associated with susceptibility to CD [366, 367]; B) *NOD2*, a gene associated with monogenic forms of CD [5], and CD susceptibility [2, 3]; C) *XIAP*, identified as a monogenic IBD gene [57]; D) *OR10H5*, a gene part of the family of olfactory receptors, known for being highly polymorphic [368].

3.3 Methods

3.3.1 Recruitment and data collection

Patients with inflammatory bowel disease were recruited through the Southampton Genetics of IBD study at Southampton General Hospital (REC: 09/H0504/125). This study has been recruiting patients since 2012, and is still recruiting patients. As such, the number of patients in the cohort is continuously growing. Both paediatric and adult IBD patients were recruited for the study in their respective clinics. Patients under the age of 18 years were diagnosed according to the modified Porto criteria [369], and adult patients were diagnosed according to the guidelines detailed in [370]. DNA for whole exome sequencing was obtained from peripheral venous blood samples collected in EDTA by the salting out method [371]. The Qubit 2.0 Flurometer was used to estimate DNA concentration, and the 260:280 ratio was calculated with a nanodrop spectrophotometer (if the DNA concentration is too low, it may not be possible to sequence the sample). Concurrently, blood for the plasma used for reactive oxygen species assays was taken. This blood was frozen until antioxidant potential and oxidative stress assays were performed. Participant's blood is usually taken at several time points as part of monitoring each patient's condition. Common laboratory tests on blood are performed, such as creatinine, albumin and C-reactive protein. Further details on the clinical data available and their extraction from University Hospital of Southampton systems are included in Chapter 5.

3.3.2 Exome data processing

Approximately 20µg of DNA was extracted from each sample and sent for WES externally. Using 1µg genomic DNA per sample, this was fragmented and enriched with Agilent SureSelect All Exon capture kit (version 4, 5 or 6). Libraries were subsequently sequenced on Illumina platforms. Samples were sequenced using paired-end sequencing, with reads varying in length depending on the sequencing batch (100, 125 and 150 base pair reads). The pipeline for exome sequencing data processing is detailed in Figure 19. After sequencing, exome data are provided as fastq files that contain the genetic sequence, where each base is paired with an ASCII (American Standard Code for Information Interchange) character that represents sequence quality (Figure 20A). Quality coded in ASCII ranges from "!" representing the worst quality, to "~" representing the best quality (Figure 20B). The quality score represents the error probability i.e. for each sequenced base, what is the probability that the base is incorrect.

The bioinformatic pipeline began with concatenation of fastq files if there were multiple lanes sequenced. The paired-end exome sequencing data was aligned to the human genome assembly

GRCh38, using the Burrow-Wheeler Aligner, BWA-MEM [80]. The default recommended parameters for mapping indels were used with BWA-MEM (open gap penalty=6, extension penalty=1). Samtools [372] was used to convert the sam (sequence alignment map) into its corresponding binary version, a BAM (binary alignment map) file. After these steps, Picard [373] sorts the BAM file by base pair coordinate (SortSam), flags duplicates (MarkDuplicates) for downstream tool GATK's HaplotypeCaller, and verifies forward and reverse strands match (FixMateInformation), correcting this if they do not. GATK 4.0 [374] then recalibrates base quality, adjusting under- and over-estimations of sequencing quality due to systematic technical errors. Firstly, the recalibration table to accomplish this is built by BaseRecalibrator, then implemented with ApplyBQSR. The recalibrated BAM file is then ready for variant calling.

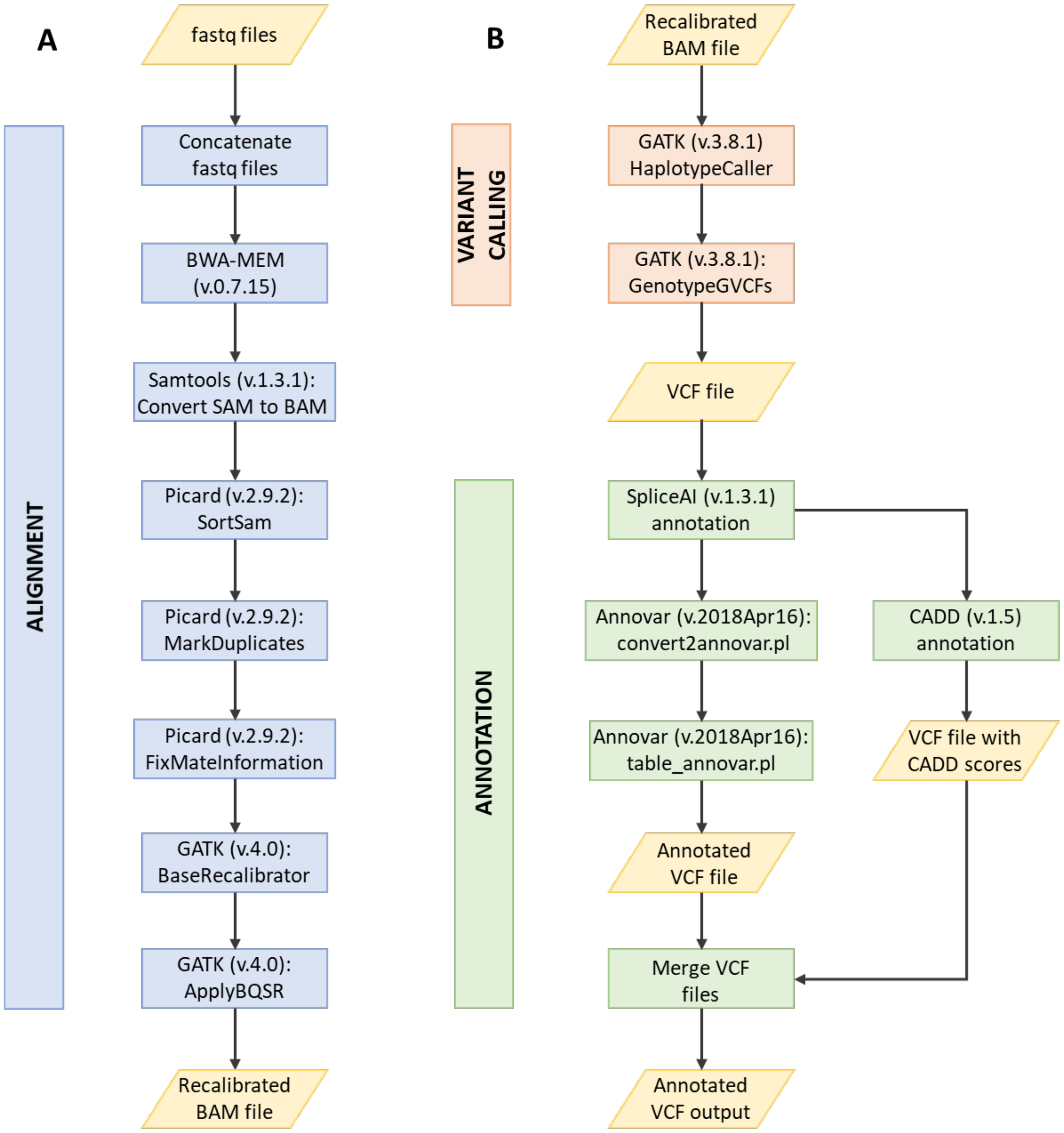
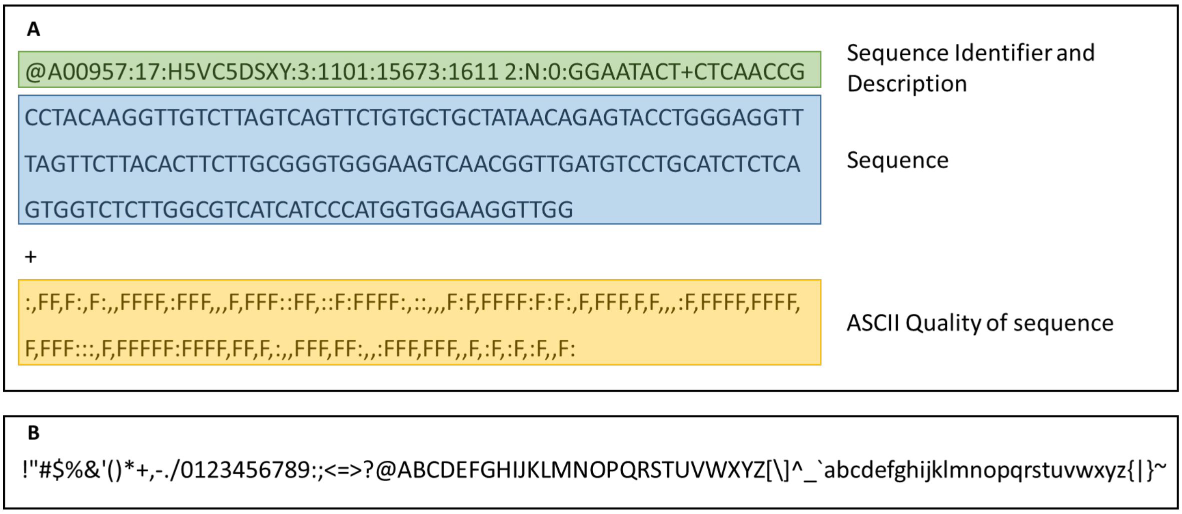

A
fastq files
ALIGNMENT
Concatenate fastq files
BWA-MEM (v.0.7.15)
Samtools (v.1.3.1): Convert SAM to BAM
Picard (v.2.9.2): SortSam
Picard (v.2.9.2): MarkDuplicates
Picard (v.2.9.2): FixMateInformation
GATK (v.4.0): BaseRecalibrator
GATK (v.4.0): ApplyBQSR
Recalibrated BAM file
B
Recalibrated BAM file
VARIANT CALLING
GATK (v.3.8.1) HaplotypeCaller
GATK (v.3.8.1): GenotypeGVCFs
VCF file
ANNOTATION
SpliceAI (v.1.3.1) annotation
Annovar (v.2018Apr16): convert2annovar.pl
CADD (v.1.5) annotation
Annovar (v.2018Apr16): table_annovar.pl
VCF file with CADD scores
Annotated VCF file
Merge VCF files
Annotated VCF output

Figure 19 The pipeline for processing an individual's exome sequencing data. This details the bioinformatic tools for each of the three main processing steps, 1) alignment to the

reference sequence; 2) variant calling to establish sequence differences between the individual and the reference sequence; and 3) annotation for variant interpretation.



```
@A00957:17:H5VC5DSXY:3:1101:15673:1611 2:N:0:GGAATACT+CTCAACCG
CCTACAAGGTTGTCTTAGTCAGTTCTGTGCTGCTATAACAGAGTACCTGGGAGGTT
TAGTTCTTACACTTCTTGCGGGTGGGAAGTCAACGGTTGATGTCCTGCATCTCTCA
GTGGTCTCTTGGCGTCATCATCCCATGGTGGAAGGTTGG
+
:,FF,F:,F:,,FFFF,:FFF,,,F,FFF::FF,::F:FFFF:,::,,,F:F,FFFF:F:F:,F,FFF,F,F,,,:F,FFFF,FFFF,
F,FFF:::,F,FFFFF:FFFF,FF,F,:,,FFF,FF:,,:FFF,FFF,,F,:F,:F,:F,,F:
```

```
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
```

A
Sequence Identifier and Description
Sequence
ASCII Quality of sequence
B

Figure 20 Fastq file format and ASCII quality. (A) Example of the fastq information format for one read. Sequence identifier and description provides information including the instrument name and run number, lane number, and if the read passed or not. (B) Lists the sequence of ASCII quality characters, with quality increasing from left to right.

During variant calling with GATK 3.8’s HaplotypeCaller [359], single nucleotide polymorphisms, inserts and deletions were identified, and in these regions the reads were re-assembled. Soft-clipped bases were not used during variant calling. Soft-clipping refers to the bases at the 5’ and 3’ ends of reads in cases where these ends have not been aligned to the reference sequence. This soft-clipping can be an indication of larger insertions and deletions in the exome sequence. By excluding these bases, some calls relating to larger indels will be missed, but it excludes many more false positive variants calls. This created an intermediate genotyped variant call format (GVCF) file with ERC (emitting reference confidence scores) GVCF formatting. This formatting results in a smaller GVCF file, as sections of the GVCF where there are no alternative alleles are condensed into non-variant blocks that represent genomic intervals. This GVCF was then input into GATK 3.8’s GenotypeGVCFs [359], where at points of variation in the GVCF, genotype likelihoods were calculated, and the variants genotyped and annotated. This creates a VCF file for annotation.

Finally, annotation provides additional information on the called variants. After converting the VCF file to an ANNOVAR [361] format (convert2annovar, ANNOVAR script), gene-based and filter-based databases were added to the file (table_annovar, ANNOVAR script). These were refGene [375], gnomAD exome v2.1.1 [82], dbnsfp35c (no annotations for synonymous variants) [86, 376] and HGMD2018 (Human Genetic Mutation Database) [91] (Table 6), adding the gene the variant resides in, allele frequencies, deleteriousness and conservation metrics, and previously reported disease(s) associated with the variant, respectively. It was also useful to include the deleteriousness metric CADD [88], but the version that annotates insertions and deletions (v.1.5) was not available using the ANNOVAR software, therefore the VCF file was annotated separately with CADD Phred scores, and merged with the annotated file. An individual's annotated file could now be filtered to find potential disease-causing variants. These annotation steps can also be completed using a VCF file that contains multiple probands, for example a cohort VCF file.

Table 6 Key databases for variant interpretation. Lists the annotation database and corresponding contents.

Database	**Contents**
refGene [375]	FASTA sequences for all annotated transcripts in RefSeq Gene
GnomAD exome v.2.1.1 [82]	Allele frequencies for all variants documented in the database. This includes the overall allele frequency of the population included in the database, as well as the allele frequency in specific subpopulations (male and female allele frequencies, allele frequencies in different ethnic groups).
dbnsfp35c [86]	Annotation of non-synonymous SNPs. This includes: • Tools that score based on whether a variant is likely to be damaging: whole-exome SIFT score, PolyPhen2 (HVAR database for Mendelian disease, HDIV for rare variants in complex disease), MutationTaster, MutationAssessor, FATHMM, PROVEAN. • Tools that score variants based on how conserved the genomic site is: GERP++, fitCons, PhyloP and SiPhy (latter two scores from previous version dbnsfp33a). • Tools that incorporate machine learning into their prediction of variant deleteriousness: CADD (v1.3), DANN, MetaSVM, MetaLR, VEST, M-CAP, fathmm-MKL, Eigen, and GenoCanyon.
HGMD2018 [91]	Published information of gene variants responsible for human inherited diseases
CADD [88]	Raw score and Phred score. The Phred score is more informative for the interpretation of variant deleteriousness.

3.3.3 Quality control

Quality control was completed on individual batches as sequencing was completed. One method by which samples can be checked for contamination is by calculating the number of shared variants between every pair of samples in each batch. A higher number of shared variants among individuals that are not known to be related indicates that either one sample has been cross-contaminated with another, or the individuals are related and this is currently unreported in the available clinical information. A lower number of shared variants could indicate contamination from another substance (not another sample's DNA). The number of shared variants between those in the batch were calculated using sample annotated VCF files, and a sample-by-sample matrix created. From the output of the script, unrelated individuals are expected to share between 60 and 65% of their variants. First degree relatives are expected to share 80-85% of their variants, and for pairs of samples where the ancestry is different for each individual it is expected that there will be fewer shared variants (approximately 55-60%). The software VerifyBamID [377] was used as an additional check for sample contamination, it outputs a p-value indicating the probability of contamination based on the sample's recalibrated BAM file.

The coverage of every sample was also checked using the BAM file. Both the mean coverage over the sample, and the read depth percentage was assessed. Each sample should have at least 20x read depth over 80% of the targeted regions. If this were not the case then the sample would have to be re-sequenced as there is not the required depth of information for downstream analyses. The annotated VCF files were used to check the percentage of heterozygous X chromosome calls, and match the sex indicated by the variant calls to the clinically recorded sex. If the percentage of heterozygous X chromosome calls were not as expected from the clinical information (55-65% in females, 10-20% in males), this could indicate bias or error in the sequencing, the sex had been misreported in the clinic, or a potential sample swap. The final check to the sequencing was using SNP fingerprinting. For every sample 24 SNPs were genotyped [356] and these genotypes were compared to the genotype called from the whole exome sequencing data. This SNP panel was designed such that it could be utilised with a number of capture kits, including Agilent capture kits [356]. This is primarily a check to ensure that the IDs of samples have not been switched, but a lack of concordance between the SNP genotypes and sequencing genotypes could also indicate contamination.

3.3.4 Joint calling and filtering for a high-quality cohort VCF file

To collate samples into one cohort VCF file, joint calling must be performed (Figure 22A). After alignment and variant calling of individual samples (as in Section 3.2.2, this was performed with the **ALIGN.sh** and the **CALL.sh** scripts), all samples must be genotyped together. By calling all samples together genotype calls are given for every site across the entire cohort, so it is possible to determine whether a site is homozygous for the reference allele, or if the data is missing (this would not be possible if sites were not called together). First, GVCFs were combined into small batches of approximately 20 files using GATK [359] CombineGVCFs (performed with **combiner.sh** script). Batches were subsequently genotyped together with GATK [359] GenotypeGVCFs, by genotyping a single chromosome for all batches, and concatenating the chromosomes with GATK CatVariants (performed with **gtyper.sh** and **catvars.sh** scripts). This creates the multi-call VCF file. GATK v.3.8.1 was used throughout to multithread this computationally intensive process without changing the pipeline [359]. At this stage of processing, the multi-call VCF file is not restricted to a BED (Browser Extensible Data) file. As a minimum, each line of a BED file contains the chromosome number, and the start and end points of a section of that chromosome. Each line can also contain additional information, such as the number of exons within the section, and their sizes. Every capture kit used prior to exome sequencing has a corresponding BED file that contains the genomic coordinates of the regions that were targeted by the capture kit. As multiple capture kit versions were used in sequencing individuals in the cohort, bedtools intersectBed was used to create a BED file of the intersection of version 4, 5 and 6 capture kits [378]. The multi-call VCF file was subsequently restricted to this BED file using GATK SelectVariants.

It is important to impose some restrictions on minimal data quality to the multi-call VCF file created in the steps above, since errors occur during sequencing, and these are particularly prevalent in genomic regions with low coverage. Retaining erroneous calls could bias outputs derived from the multi-call VCF file, and downstream analyses. To improve the quality of the multi-call VCF, the methods described by Carson et al. [379] were implemented (Figure 22B). The genotypes of variants with a sequencing depth less than 8 and genotype quality (GQ, confidence that the genotype is correct) less than 20 are replaced with the missing genotype (./.) using vcftools [380], as these are poor quality variants. A filter requiring the mean GQ for each variant across all included samples to be greater than 35 was also applied, followed by a missingness filter to ensure that each variant was genotyped in a minimum of 88% of the samples in the VCF, achieved using vcftools [380]. These filtering steps were performed using **Filtering.sh**.

New quality scores based on the likelihood of a variant being true versus being a sequencing artefact were calculated by GATKs [359] VariantRecalibrator, which uses a Gaussian mixture model to evaluate each variant. The model parameters created are applied using ApplyRecalibration, which annotates the files with new quality scores, and flags those that do not meet the required quality threshold (Figure 21). These two steps combined complete the Variant Quality Score Recalibration (VQSR) process. In this case, the quality threshold was tranche 0.99, which requires 99% of the variants in the VCF to be included. This process was performed twice so that SNVs and Indels were evaluated separately. In the case of the Gaussian mixture model for Indels there were fewer variants, so the maximum number of Gaussians was set to four. This lowers the number of clusters in the Gaussian mixture model so that there are enough variants per cluster to satisfy modelling requirements, but this comes at the expense of resolution, i.e. reduced ability to identify sequencing artefacts (recalibration stages performed using **Recalibrate.sh**). Vcftools [380]was used to remove the flagged variants, creating a high-quality multi-call VCF that can be used to create GenePy scores.

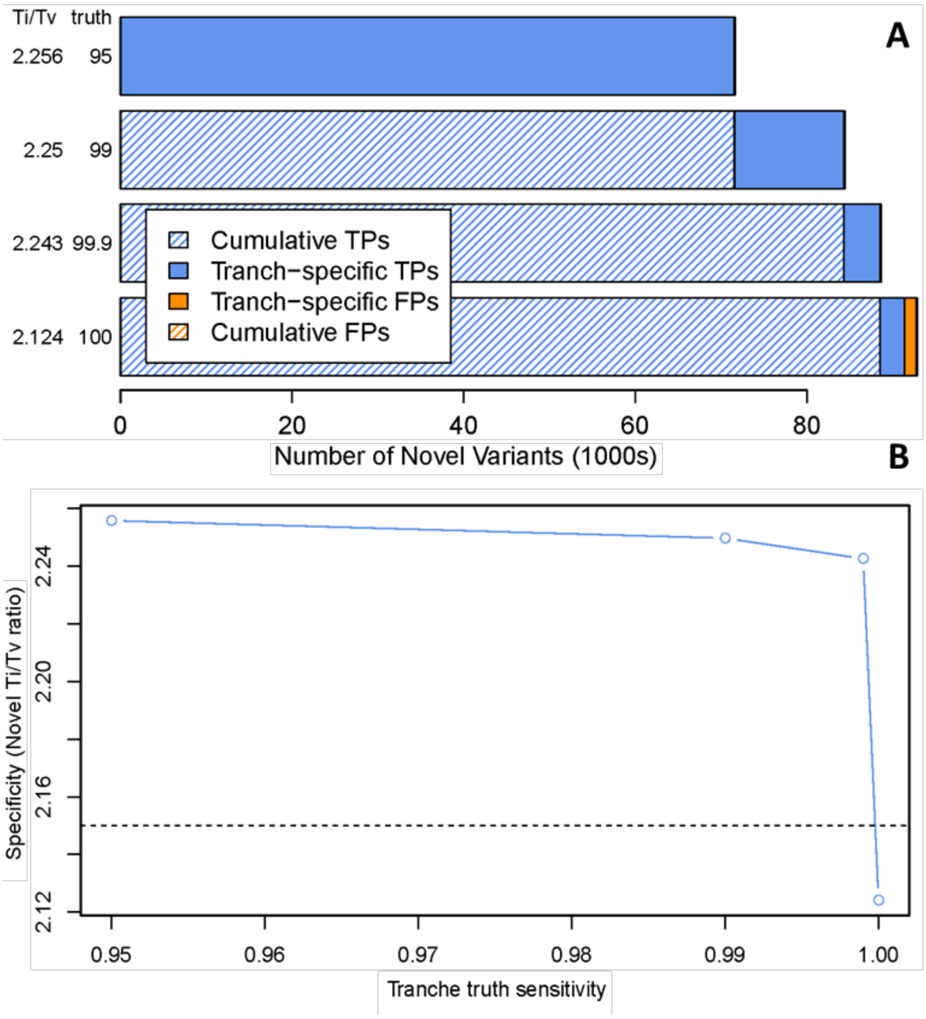


Figure 21 Example of plots generated by running GATKs Variant Recalibration and Apply Recalibration for a cohort (n=491). The transition/transversion ratio is a good proxy for the true positive/false positive trade-off. Transitions are a base changing to another base that has the same chemical structure, transversions are a base changing to another base with a different chemical structure. (A) Breakdown of true positives and false positives per tranche. Inclusion of all variants (tranche 100) would lead to false positive variants (assigned the label false positive by the Gaussian Mixture Model) remaining in the call set. (B) Specificity decreases (novel

transition/transversion ratio) as the sensitivity (tranche) increases. The tranche is chosen to optimise the trade-off between these two metrics.

3.3.5 Annotation and generation of GenePy matrix

The algorithm that creates GenePy scores requires that only bi-allelic variants are present in the VCF, so these were removed with vcftools [380]. Only using bi-allelic variants is a limitation, caused primarily by data availability. To include tri-allelic and quad-allelic variants another database would need to be used to obtain allele frequencies, as the gnomAD database only reports the major alternative allele frequency [82]. The genome browser Ensembl [381] reports multiple alternative alleles, but integrating this primarily online data source into current pipelines is difficult. Most importantly, Ensembl's data is not as complete as data from the gnomAD database. If future updates to gnomAD include reporting of minor alternative alleles, GenePy could be modified to include all variants.

ANNOVAR [361] was used to create an annotated VCF file that included the gene name for each variant (refGene database [375]) and the allele frequency in the general population (gnomad_exome, all individuals [82]). If the variant is novel to the gnomAD database, it will be assigned a frequency of 1/282,912 for the purposes of calculating a GenePy score. The denominator is twice the number of exomes and genomes in the gnomAD database, because there are two opportunities (i.e. two alleles) for the variant to appear in every person included in the database. ANNOVAR does not currently support the most recent version of CADD, so the CADD scores for each variant were generated separately. Required columns from the annotated VCF and the file containing CADD [88] scores were merged together (merging performed with **cross-annotate-cadd.py**). The file was then filtered to retain the type of variant required, for example exonic and splicing. Subsequently the GenePy script was run for the genes that scores were required for, either for all available genes from the RefSeq database [375], or a specific list. This script creates files with GenePy scores for the cohort for each gene that are subsequently merged into a matrix containing one score per patient per gene. Steps for creating the GenePy matrix are summarised in Figure 22C. Several scripts were utilised to generate the GenePy scores: **subber.sh**, **GenePy_1.3.sh**, **make scores_mat_6.py**, **generate_final_matrix.py**, and **MatrixMaker.sh**. Full instructions for generating this matrix are found in the Supplementary Files.

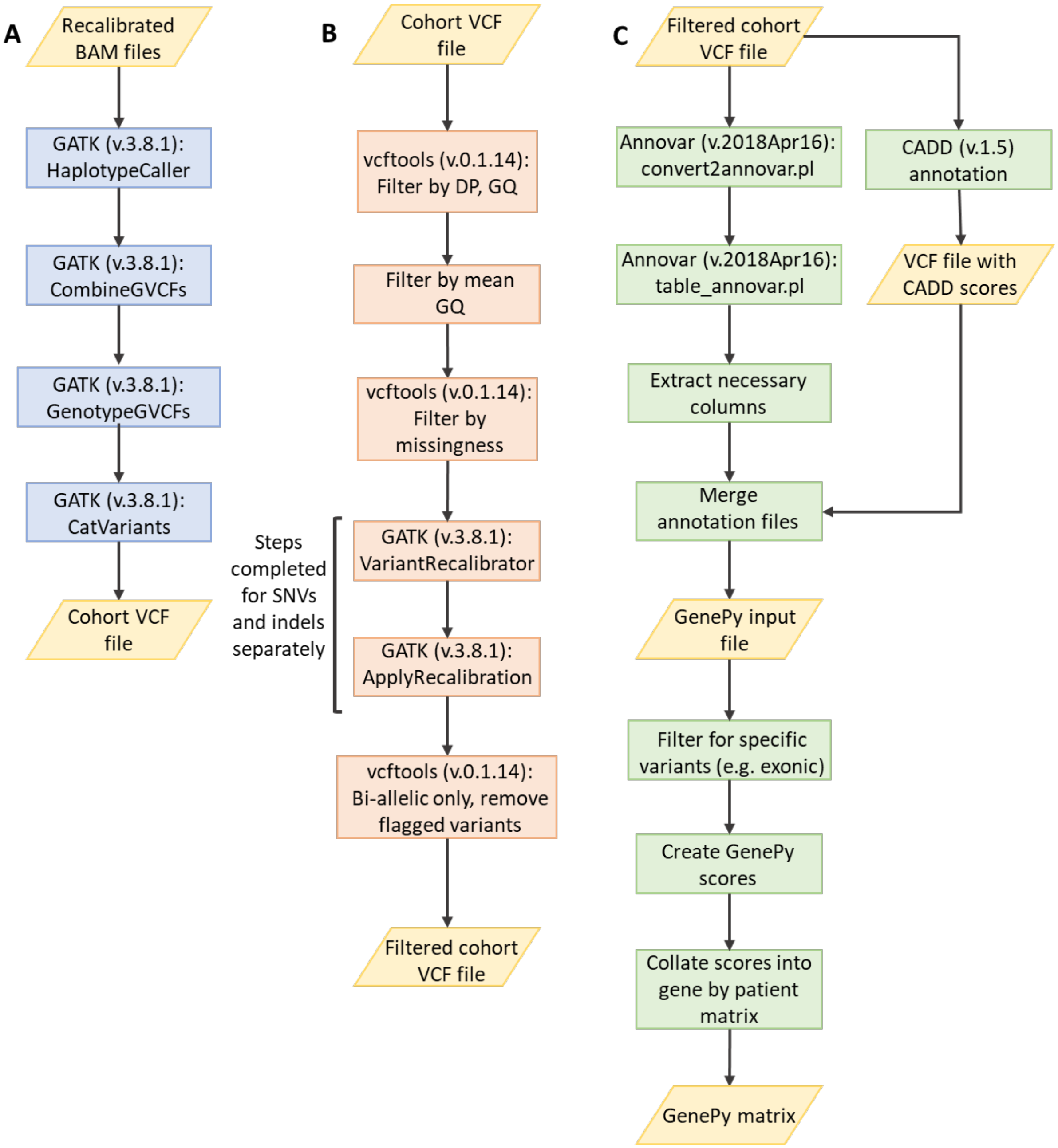

A
Recalibrated BAM files
GATK (v.3.8.1): HaplotypeCaller
GATK (v.3.8.1): CombineGVCFs
GATK (v.3.8.1): GenotypeGVCFs
GATK (v.3.8.1): CatVariants
Cohort VCF file
B
Cohort VCF file
vcftools (v.0.1.14): Filter by DP, GQ
Filter by mean GQ
vcftools (v.0.1.14): Filter by missingness
Steps completed for SNVs and indels separately
GATK (v.3.8.1): VariantRecalibrator
GATK (v.3.8.1): ApplyRecalibration
vcftools (v.0.1.14): Bi-allelic only, remove flagged variants
Filtered cohort VCF file
C
Filtered cohort VCF file
Annovar (v.2018Apr16): convert2annovar.pl
Annovar (v.2018Apr16): table_annovar.pl
Extract necessary columns
CADD (v.1.5) annotation
VCF file with CADD scores
Merge annotation files
GenePy input file
Filter for specific variants (e.g. exonic)
Create GenePy scores
Collate scores into gene by patient matrix
GenePy matrix

Figure 22 Three stages for GenePy matrix creation. (A) Variant calling and genotyping samples together to create a multi-call VCF file. (B) Filtering and recalibration to improve the quality of the cohort VCF. (C) ANNOVAR and CADD annotation to provide necessary information for GenePy scoring (frequency, deleteriousness), subsequent creation of GenePy scores and collation into a patient by gene matrix.

3.4 Pipeline developments for cohort analysis

In 2020, WES data became available for a large batch of adults with IBD recruited as part of the Genomics of IBD study. The sequencing of this batch was performed as part of the National Institute of Health Research’s BioResource [382].This led to the number of individuals for which WES data was available approximately doubling, as previously fewer than 500 paediatric patients with WES data were present in the cohort. Additionally, all paediatric DNA samples that had been previously sequenced with version 4 of the Agilent SureSelect Human All Exon capture kit were re-sequenced using version 6 of this capture kit. Substantial discrepancies existed between the exon coverage of version 4, and versions 5 and 6 (Figure 23). This was therefore a desirable improvement to the exome capture efficiency. These changes resulted in a larger and more uniform dataset. Therefore all whole exome sequencing in the cohort had been performed with SureSelect version 5 or 6 capture kits. While this increase in data was welcome, it was also recognised that this would increase the time required to process the data significantly. The decision was made to update the exome processing pipeline to utilise the latest version of GATK (v.4.1.2) during the joint calling process (GATK v.3.8 was previously used), utilising a new joint calling workflow that would uplift called variants [383, 384]. The annotation process was also revamped by replacing ANNOVAR [361]with Ensembl-VEP [362]. The new pipeline, intended to align, joint call and annotate to enable the transformation of WES data into a cohort GenePy matrix, is described in the next sections.

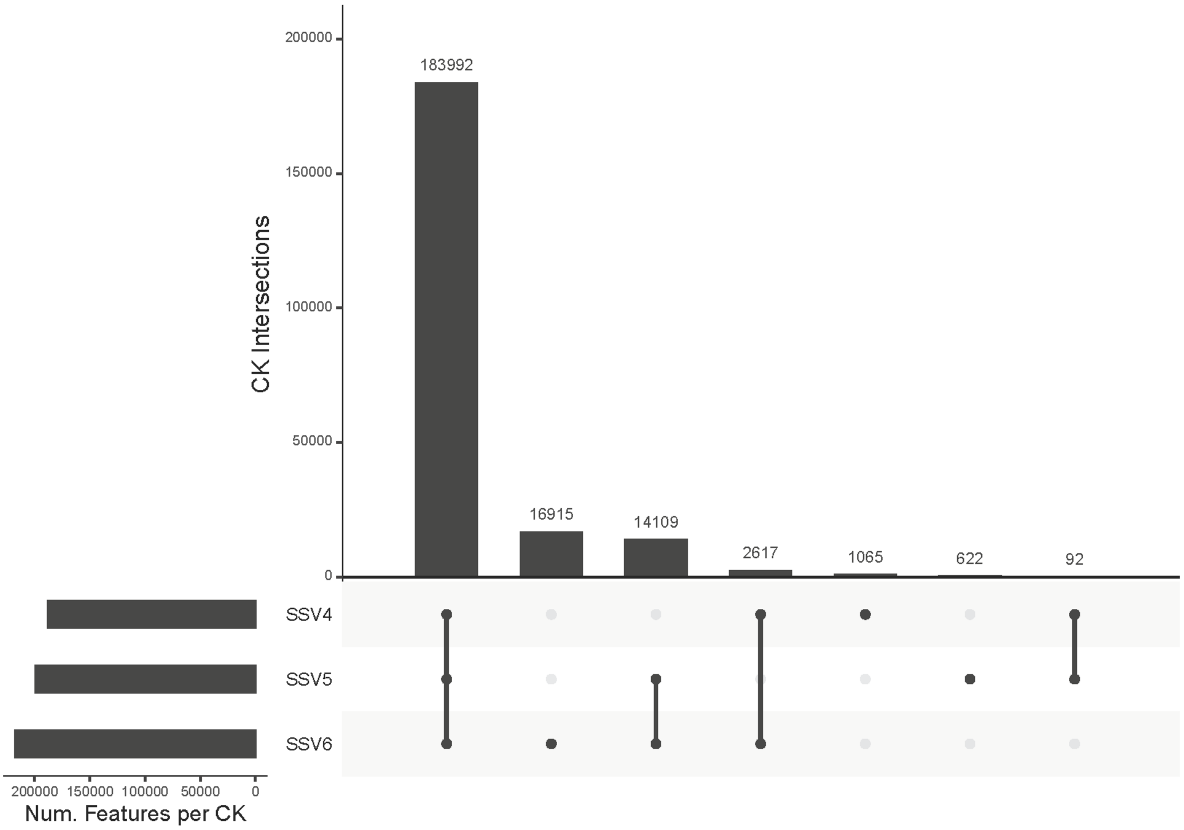


Figure 23 Intersection of features in SureSelect version 4 (SSV4), version 5 (SSV5) and version 6 (SSV6) capture kit (CK). Feature counts determined with Python's Pybedtools [385], plotted using R's UpSetR [386]. Plot shows a large overlap between all three capture kits, but there are more features in the BED file that are 1) unique to SSV6; and 2) present in SSV5 and SSV6 that are absent in SSV4, demonstrating the desirability of resequencing where this was performed with SSV4.

3.4.1 Alignment and joint calling

Alignment proceeded as detailed within Section 3.3.2 and Figure 19A, with two differences. Firstly, for the new pipeline an updated GRCh38 reference sequence was employed which included HLA decoy sequences. Secondly, for the use of GATK v.4.1.2 [383, 384]in downstream joint calling, there was a requirement to use a specific version of Java, OpenJDK, for generating a BAM file from the fastq files. The GATK version utilise for BaseRecalibrator and ApplyBSQR had already been updated from v.3 to v.4, so this portion of the script remained the same (alignment was achieved with the **preprocess.sh** script).

To begin the joint calling process, first each BAM file created during alignment had to be individually called. This was achieved as described in Section 3.3.2 and Figure 19B, except with the GATK version 4.1.2 [383, 384]. Briefly, HaplotypeCaller converts the BAM file into a GVCF file, and then these files are converted to VCF with GenotypeGVCFs. Additionally, an interval list was given to HaplotypeCaller in order to speed up joint calling downstream. The interval list was based on the union of the SureSelect V5 and V6 capture kits, with 150bp padding on these intervals (individual calling accomplished with **caller.sh**).

The main change to variant calling is in the joint calling stage (performed with **joint_calling.sh** script). Previously, CombineGVCFs was utilised to transform many VCF files into a single file. With the new GATK version 4 [383, 384], GenomicsDB performs the same function using a different method, which improves on the time required to joint call many samples. GenomicsDB is essentially a database for the variant call information. This data is stored in a 2D TileDB array, where the rows represent a sample in the cohort, and the columns represent a genomic position, which includes chromosome and base position. Therefore, each cell contains that sample's information at that genomic position. GenomicsDB can take many arguments which means joint calling can be customised according to the user's processing power. It also has an important variation on the GenomicsDB command, called GenomicsDBImport, which allows samples to be added to an existing database. This enables samples to be read into the database in batches. For joint calling this cohort, GenomicsDBImport was used, and batches of 30 samples were read in to the database. It is also required to provide a list mapping every sample name to its file location. The previously described interval list was split into 97 smaller interval lists using Picard, so that a different database for each of these regions, for all samples, was generated. In order to speed up this process, the option to import data between the intervals was used (merge-input-intervals), as this is recommended for WES data, where there are many intervals. Multithreading was used for opening batches of VCF files, also to improve the processing time.

These 97 genomics databases, each containing all samples, were then joint called with GATK v4 [383, 384] GenotypeGVCFs. Performing the joint calling in this way is an improvement on doing so in small batches of 30, as was done previously. This is because when joint calling is performed on increasingly large numbers of samples, there is an increased opportunity to identify genotypes where there is low confidence in a variant in one sample but many other samples have the variant with high confidence, which confirms the likelihood of a variant in that location. Here, joint calling has increased from 30 samples at a time in the previous pipeline, to over 1000, resulting in an increased opportunity to identify variants. For joint calling, the associated interval list is again provided. The option to load data in between intervals is included again, but this is combined with the option to only output calls that start in the given intervals. These two options together allow the process to be quicker, while restricting the intervals and so reducing file sizes. In addition, common sites from dbSNP v.151 are provided to the caller, which provides sequence quality calibration for these common alleles, and common alleles are called more reliably if this resource is provided. Picard's [373] MergeVcfs was then used to combine the 97 joint called VCF files.

3.4.2 Filtering, annotation, and GenePy matrix construction

After generating the joint called VCF file, VQSR was performed as before, utilising GATK v.4.1.2's [383, 384] VariantRecalibrator and ApplyRecalibration for SNVs and indels separately. Previously, the maximum number of Gaussians had to be set to 4 for indels to reduce the model's resolution. Due to the increase in sample size, it was no longer necessary to reduce model resolution, and therefore VariantRecalibrator's ability to identify sequencing artefacts was improved compared to previous VCF file processing. The other change to the workflow for VQSR was utilising the option called "trust-all-polymorphic" for both stages, and SNV and indel models. This option assumes that all variants are polymorphic, in other words that one or more alternative alleles are present at each site. This option improves processing time considerably, according to GATK's documentation. In this pipeline, performing VQSR prior to other quality-based filtration is a deviation from the workflow set out by Carson et al. [379]. A comparison of the number of variants present when performing VQSR before or after the other quality filters showed that the order of these steps did not impact the number of variants (these recalibration steps were performed with the **vqsr.sh** script).

Next, the VCF was restricted to the BED file of the intersection of the Agilent SureSelect V5 and V6 capture kits [378]. The file was then filtered using VCFtools [380] with all quality thresholds related to depth, GQ, mean GQ and missingness previously outlined in Section 3.3.4. As before, the VCF file was restricted to biallelic variants only. Then, the VCF file with only biallelic variants was annotated using Ensembl-VEP. This was a computationally intensive process and thus the file was split into chunks by chromosome, before annotation. The longest chromosome, chromosome 1, took approximately 24 hours to annotate. Ensembl-VEP [362] became a more desirable annotator than ANNOVAR [361] for two reasons. First, it is maintained better than ANNOVAR, as it has frequent version updates, and these are accompanied by updates to the latest versions of *in silico* tools alongside adding new tools. Secondly, Ensembl-VEP has CADD v1.6 [363] as a Plugin, enabling easier VCF file annotation with this important metric. In addition to annotating the cohort VCF file with CADD v.1.6 with Ensembl-VEP (v.103), annotations from the gnomAD v.2.1.1 database were included. Instead of RefSeq, Ensembl-VEP utilises its own sequence database. The annotator was run with the option pick allele, which means each allele will be annotated with information associated with only one gene transcript. The default order for choosing the transcript was used, where the canonical transcript was utilised if it was available (annotation performed with the **vep.sh** and **vep_x.sh** scripts). After annotation, the individual chromosome VCF files were concatenated together into one VCF file.

Separately, the input VCF was annotated by ANNOVAR [361] with the gnomAD random forest flag, from gnomAD v.3.1.1. This method was designed by the Broad Institute as a flag-based quality filter for their gnomAD database [82]. Briefly, a random forest machine learning model takes in information from resources such as the 1000 Genomes high-quality site dataset, and classifies variants as being true polymorphisms, or sequencing artefacts. This sequencing artefact category is additionally broken down into other categories based on the variant characteristic(s) that flagged it as dataset noise. This flag can be one, or a combination of AC0; AS_VQSR, and InbreedingCoeff. Variants classified as AC0 had no alleles after filtering out low-quality calls (based on depth, GQ and allele balance); AS_VQSR variants failed allele-specific GATK VQSR, and the InbreedingCoeff flag indicates excess heterozygosity at the variant site.

Although using Ensembl-VEP [362] to annotate CADD [363]scores improved processing time, the annotator was unfortunately unable to annotate all sites. To fill in the annotation gaps, these sites were extracted from the annotated VCF file, and uploaded to the CADD website. The scores from the CADD website were re-inserted into the original data using a Python (v.3.7) script (**genepy_combine_annotations.py**). Next, all relevant columns were compiled together before final filtering: chromosome, variant start position, reference allele, alternative allele, variant consequence, gene symbol, gnomAD allele frequency, CADD Rawscore, and gnomAD random forest flag. Variants that had failed the random forest flag were excluded from the file. Finally, only exonic variants were retained for the generation of a GenePy score matrix. As Ensembl-VEP's variant consequence field was more granular than ANNOVAR, the following variants were defined as exonic:

- Coding sequence variant
- Downstream gene variant
- Frameshift variant
- Inframe deletion
- Inframe insertion
- Missense variant
- Protein altering variant
- Splice acceptor variant
- Splice donor variant
- Start loss
- Stop gained
- Stop lost
- Stop retained variant
- Synonymous variant

- Upstream gene variant

It should be noted that some variants can be annotated with multiple consequences, but as long as one of the consequences listed above was present, that variant would be included in GenePy scoring. After this the GenePy matrix was generated. The matrix was generated with the following scripts: **subber.sh**, **GenePy_1.3.sh**, **make scores_mat_6.py**, **generate_final_matrix.py**, and **MatrixMaker.sh.** More information on specific file manipulation required for VCF file preparation for GenePy, and the usage of scripts in matrix generation is detailed in the research group's Github, and a static version of these instructions is available in the Supplementary Files. The new pipeline is shown in Figure 24.

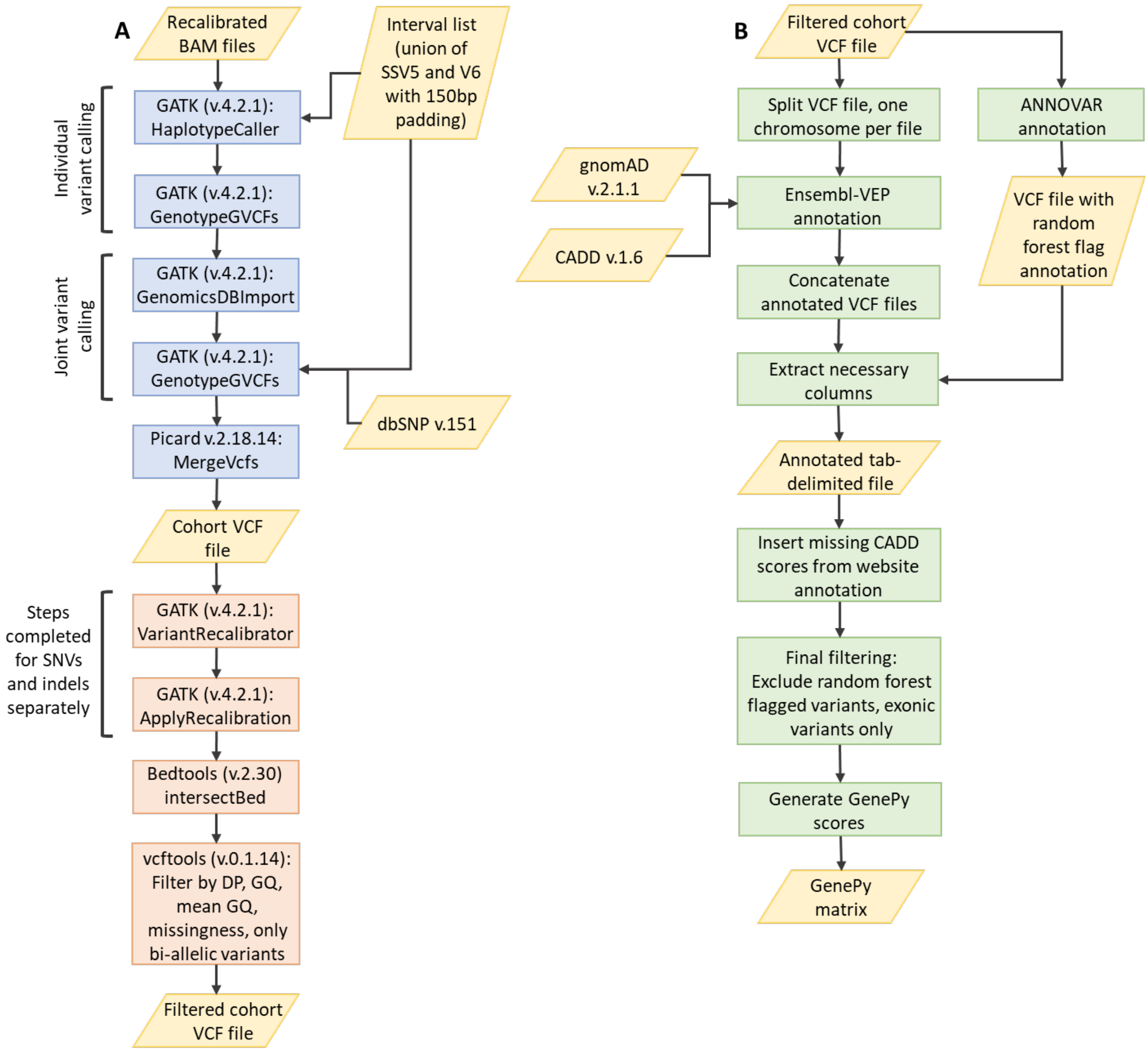

A
Recalibrated BAM files
Interval list (union of SSV5 and V6 with 150bp padding)
Individual variant calling
GATK (v.4.2.1): HaplotypeCaller
GATK (v.4.2.1): GenotypeGVCFs
Joint variant calling
GATK (v.4.2.1): GenomicsDBImport
GATK (v.4.2.1): GenotypeGVCFs
dbSNP v.151
Picard v.2.18.14: MergeVcfs
Cohort VCF file
Steps completed for SNVs and indels separately
GATK (v.4.2.1): VariantRecalibrator
GATK (v.4.2.1): ApplyRecalibration
Bedtools (v.2.30) intersectBed
vcftools (v.0.1.14): Filter by DP, GQ, mean GQ, missingness, only bi-allelic variants
Filtered cohort VCF file
B
Filtered cohort VCF file
Split VCF file, one chromosome per file
ANNOVAR annotation
gnomAD v.2.1.1
CADD v.1.6
Ensembl-VEP annotation
VCF file with random forest flag annotation
Concatenate annotated VCF files
Extract necessary columns
Annotated tab-delimited file
Insert missing CADD scores from website annotation
Final filtering: Exclude random forest flagged variants, exonic variants only
Generate GenePy scores
GenePy matrix

Figure 24 Pipeline for updated joint calling and annotation for the generation of GenePy scores. A) Variant calling and quality control. This pipeline starts at the variant calling stage, as the alignment stage is already detailed in Figure 19. Variants are called for every individual, and then joint called. Subsequent filtering steps lead to a high-quality cohort VCF file. B) Annotation and GenePy score generation. Ensembl-VEP and ANNOVAR annotate the cohort VCF file, and then file manipulations are performed in order to ensure relevant information is complete and columns ordered. Final filtering based on annotated variant quality and variant consequence is the last step before generating the GenePy scores.

3.5 Research outputs from the processing of whole exome sequencing data

Throughout my candidature, exome data was processed and multiple GenePy matrices generated as more patients were recruited to the Genetics of IBD study. In addition to the use of GenePy matrices throughout this thesis, this data processing also contributed significantly to other research outputs.

In a study led by James Ashton, I assisted with the processing of WES data. This research focussed on variant-level data, and sought to identify paediatric patients within the IBD cohort where their disease could be considered to have a monogenic cause [5]. This analysis led to the discovery of patients with “pathogenic” or “likely pathogenic” variants (according to American College of Medical Genetics guidelines) in several genes, including *TRIM22*, *WAS*, and *NOD2*. Often, these patients had variants which were compound heterozygous, and this was later confirmed through segregation analysis. Additionally, patients thought to have an autosomal recessive *NOD2*-related disease were more likely to have a stricturing (narrowing of the gastrointestinal tract) phenotype.

For another study that utilised genomic data and targeted RNA-sequencing data obtained from ileal biopsies in treatment-naïve paediatric patients, I processed WES data, and generated GenePy matrices. This research found that high GenePy scores in several genes across the NOD-signalling pathway, including NOD2, were associated with reduced transcription of the NF-κB pathway [387]. Another study led by James Ashton used the GenePy scores of *NOD2* generated by myself to attempt to stratify patients into several risk groups based on the presence or absence of the CD stricturing endotype (narrowing of the gastrointestinal tract) [388]. The presence of the stricturing endotype in the highest risk group was over 50%, compared to approximately 20% of patients with a stricturing endotype in the lowest risk group.

A GenePy matrix I generated was also used for a study by Enrico Mossotto and Joanna Boberska, which also utilised metabolomic data [389]. They sought to establish links between paediatric patient’s active inflammation, their metabonomic profiles, and their genomic variation in the form of GenePy scores. This was accomplished by using machine learning to identify key nuclear magnetic resonance (NMR) peaks, correlating these peaks with GenePy scores, and finally performing gene enrichment analysis.

I provided WES data, or GenePy matrices for two studies led by Tracy Coelho. One study focused on periostin, a matricellular protein implicated in tissue fibrosis, and its potential use for assessing disease activity and surgical outcomes. As such the GenePy scores of a gene network functionally connected to periostin was examined. They found no significant differences in mutational burden in these periostin-connected genes between patients who did and did not undergo surgery [390]. The second study, which researched the immunological profile of paediatric IBD patients, looked at muramyl di-peptide (a peptidoglycan motif on several types of Gram-positive and Gram-negative bacteria)immune response in the context of *NOD2* variants [391].

Chapter 4 Reactive oxygen species and inflammatory bowel disease

Chapter summary – in this chapter, data from oxidative stress and antioxidant potential assays are analysed. The relationship between these assays and patient characteristics were assessed. This included observing any relationship between these assays and age, IBD diagnosis, and the commonly used blood marker for general inflammation C-reactive protein. Additionally, there was an interest in understanding how the patient's assay results related to their genetics. This was assessed by using GenePy scores of genes involved in reactive oxygen species pathways, combined with linear regression, and then machine learning. The machine learning pipeline tested several different algorithms to find the method that could best differentiate between high and low assay results (for each assay), according to the GenePy scores for key selected genes.

Chapter contributions – the FRAP (ferric reducing ability of plasma), TBARS (thiobarbituric acid reactive substances), and TFT (total free thiol) assays were performed by Magda Minnion, Bernadette Fernandez, and Monika Mikus-Lelinska. Martin Feelisch assisted with the interpretation of raw data from these assays. Enrico Mossotto assisted with the processing of WES data. All analysis performed by Imogen Stafford.

Supplementary files can be found at https://doi.org/10.5258/SOTON/D2655.

4.1 Introduction

As discussed in Section 1.2.6.3, the production of reactive oxygen species (ROS) is of interest to IBD research due to dysregulation of their production potentially leading to intestinal inflammation [103, 110]. Although ROS are important signalling molecules (redox signalling) for downstream immunological pathways, excessive ROS production can cause cellular oxidative stress. Oxidative stress can cause oxidative damage to biomolecules including lipids, protein and DNA. To counteract excessive ROS production, antioxidant molecules can be released that scavenge ROS in order to prevent oxidation of molecules, decreasing oxidative stress [392]. In order to gain insight into the impact of ROS in disease, both oxidative stress and the ability, or potential, to produce antioxidants must be measured.

4.1.1 Assays for oxidative stress and antioxidant potential

Just as the blood marker C-reactive protein (CRP) is used as an indication of inflammation levels in a patient, measurements of oxidative stress in patients could also be indicative of inflammation. Markers of inflammation can indicate whether disease is active or in remission. The ferric reducing ability of plasma (FRAP) assay measures the total antioxidant capacity, and thus can measure its capability to handle future oxidative stress events. The thiobarbituric acid reactive substances (TBARS) assay measures the oxidative degradation of lipids (lipid peroxidation) by ROS. Lipid peroxides are highly unstable, and their metabolism generates TBARS, including malondialdehyde. The Total free thiol (TFT) assay is a more general measure of oxidative stress that assesses the total thiol status of the plasma, and describes the plasma antioxidant status in the body. Through measuring the thiol antioxidants, the assay is a proxy for current oxidative stress. These assays are all performed using plasma, which makes using them as a marker relatively accessible in clinical settings because blood is all that is required from the patient.

4.1.2 Measuring oxidative stress and antioxidant potential in autoimmune disease

Reactive oxygen species are known to be involved in chronic granulomatous disease, but are also implicated in the pathogenesis of a number of autoimmune diseases. Mateen et al. confirmed an increase in production of ROS along with increased lipid peroxidation in rheumatoid arthritis (RA) in comparison to healthy controls. In comparison to controls RA patients also had a reduced capacity to defend against oxidative stress with antioxidant production (FRAP assay) [393]. Lower antioxidant potential has also been observed in patients with juvenile idiopathic arthritis in comparison to healthy controls, although the difference only trended towards significance [394]. This decreased antioxidant potential in comparison to controls was also found in another group of children diagnosed with type 1 diabetes [395]. Juybari et al. also observed significantly higher lipid peroxidation and reduced antioxidant capacity in relapsing remitting multiple sclerosis patients in comparison to healthy controls [396]. Significantly higher lipid peroxidation in comparison to healthy controls has also been identified in coeliac disease [397]. These case-control studies do provide more evidence of ROS involvement in these diseases, but do not progress the question of whether they can be used in clinical management. There are fewer studies that examine the heterogeneity of ROS production and antioxidant response within autoimmune diseases and how this relates to disease course. Ademoglu et al.'s study of Grave's disease patients established differences in lipid peroxidation depending on disease course. Patients who experienced a relapse after treatment had significantly higher levels of lipid peroxidation in comparison to patients in remission [398].

It is known that variation in NADPH oxidase genes is directly linked to chronic granulomatous disease (CGD) aetiology. The disease results in patient's susceptibility to severe bacterial and fungal infections, but disease can also manifest with a number of inflammatory conditions, including intestinal inflammation phenotypically similar to CD. This susceptibility to infection is caused by deficiencies in NADPH oxidase that results in decreased production in ROS. This appears to be at odds with the apparent conclusion that the contribution of ROS to autoimmune disease is through overproduction of these molecules. However, it is now thought that a hyper inflammatory immune response is due to autophagy dysregulation caused by ROS deficiency. ROS are necessary to facilitate autophagy, and if this does not occur then there is an increased production of interleukin 1β, an inflammatory cytokine [399]. This cytokine is known to contribute to IBD pathogenesis, contributing to CD-like disease in CGD patients. Currently, there is very limited research into lipid peroxidation and antioxidant capacity in CGD. It is therefore unknown whether the same trends in increased lipid peroxidation and decreased antioxidant capacity (in comparison to healthy controls) in other autoimmune diseases would be observed in CGD patients. This type of research could provide some evidence as to whether underlying ROS mechanisms in CGD are similar or different to other autoimmune diseases such as rheumatoid arthritis, multiple sclerosis and type 1 diabetes.

4.1.3 Measuring oxidative stress and antioxidant potential in inflammatory bowel disease

Of the three assays described in Section 4.1.1, the majority of prior research has used either the FRAP or TBARS assays, with little evidence that the TFT assay has ever been used for an IBD cohort. The FRAP assay has been used to evaluate differences in antioxidant status of both serum and saliva in those with active and inactive CD. The CD group with active disease had decreased antioxidant capacity in comparison to patients with inactive CD, where the measure of activity was the CD activity index [400]. Szczeklik et al. also investigated serum and saliva antioxidant potential differences between CD and UC, finding a significantly lower antioxidant capacity in CD patients [401]. Both studies were conducted with smaller cohorts of 58 and 31 IBD patients, respectively. A study by Luceri et al. used both the FRAP and TBARS assays to evaluate oxidative stress and antioxidant potential in the serum of adults patients with severe CD requiring surgery. In this, they identified significantly higher lipid peroxidation in the CD group in comparison to controls, but no differences in the FRAP assay. This was also a relatively small study, with 54 CD patients and 17 controls [402]. Statistical significance between CD and controls in TBARS assays was corroborated by Langenberg et al. [403]. Maor et al. also confirmed significantly higher lipid peroxidation levels in CD than controls, and also in those with active disease in comparison to inactive disease, according to the CD activity index [404]. Levels of lipid peroxidation levels in UC have not been shown to be significantly higher than controls [405]. These studies in small groups of patients all show the same trend of increased lipid peroxidation and decreased antioxidant capacity as observed in other autoimmune diseases. No study analysed an exclusively paediatric IBD cohort, including few, if any, paediatric cases. As stated earlier, there were differences in results from juvenile idiopathic arthritis patients in comparison to an RA study. This could be due to the patient's ages, or differences in disease aetiology.

There are currently no biomarkers that are specific to IBD or any subtype, either for diagnosis or monitoring disease activity. As previously mentioned, CRP levels can be used as an indication of disease status in the patient. Other nonspecific blood markers, such as erythrocyte sedimentation rate, platelet count and mean platelet volume can be used as an indicator of whether disease is in remission or active. These methods are only used to guide further investigation, and cannot be used diagnostically, or to confirm remission. For subtype diagnosis, the best biomarkers are thought to be a combination of anti-Saccharomyces cervisiae antibodies (ASCA) and atypical perinuclear antineutrophil cytoplasmic antibodies (pANCA). Used together they had a specificity above 90% and sensitivity of approximately 55% when differentiating UC from CD, however sensitivity is much lower for colonic UC versus CD (approximately 35%) [406]. Faecal calprotectin is a marker currently in use for differentiating between IBD and irritable bowel syndrome, and for indicating intestinal inflammation [407].

The aim of the following research was to better understand the mechanisms through which ROS production affects IBD using the FRAP, TBARS and TFT assays. Additionally to assess whether oxidative stress or antioxidant potential were viable markers of IBD or its subtypes. To this end the relationships between: 1) the antioxidant potential and oxidative stress assay data; 2) the three assays and clinical and demographic characteristics of paediatric IBD patients, and 3) the three assays and genomic data converted into GenePy scores were explored.

4.2 Methods

4.2.1 Genomic data

Paediatric patients were recruited and blood for DNA and plasma collected as described in Section 3.3.1. WES data processing, including alignment, joint calling, annotation and GenePy score generation was performed as in Section 3.3.

4.2.2 Antioxidant potential and oxidative stress assays

The metabolomic assays analysing oxidative stress and antioxidant potential were performed over three days, across 11 (FRAP and TBARS assays) or 17 (TFT assay) plates, using previously frozen plasma samples. All three assays are based on spectrophotometric methods and performed in duplicate (FRAP and TBARS assays) or triplicate (TFT assay and protein content).

4.2.3 Ferric Reducing Ability of Plasma assay

The protocol for the FRAP assay is based on the method by Benzie and Stain [408]. In brief, the FRAP reagent was prepared, which mixes 25 ml acetate buffer, 2.5ml tripyridyltriazine solution, and 2.5ml iron (III) chloride solution. Calibration was performed with iron (II) solutions of known concentrations between 100 and 1000 µmol/L. 300 µL of the FRAP reagent was incubated at 37^{0}C for 30 minutes. After a reagent blank reading was taken with the FRAP reagent at 593nm on a spectrofluorometer, 10µL of centrifuged plasma, which was thawed as required, was added to the FRAP reagent alongside 30 µL of H_2O. The absorbance of the resulting blue colour from the reduction of ferric ions to ferrous ions measured at 593nm on a spectrofluorometer.

4.2.4 Thiobarbituric acid Reactive Substances assay

The TBARS assay usually requires a larger volume of plasma, but due to plasma availability this assay was miniaturised [409]. In brief, the TBARS reagent was prepared by dissolving 7.5g trichloroacetic acid, 1.035ml hydrochloric acid, and 0.1875g 2-thiobarbituric acid in 50ml of milliQ-water. Additionally the Butylated hydroxytoluene (BHT) solution was prepared by dissolving 0.2g of BHT in 10 ml ethanol. Next, 65µL of plasma and 65µL of methanol were centrifuged, and the recovered volume (approximately 90µL) was added to a 1:1 mix of TBARS and BHT reagents. This mixture was transferred to glass inserts inside microcentrifuge tubes with 500 µL water inside, and incubated at 90^{0}C for half an hour. At this high temperature the malondialdehyde-thiobarbituric acid adduct forms. Stainless steel balls covered the glass inserts to prevent sample evaporation, but allow the escape of excess gas. After cooling on ice for 10 minutes, and being centrifuged for 15 minutes, 50µL of each of the samples was added to microplate wells, and the fluorescence intensity read at 532nm and 750nm on a spectrofluorometer (the latter reading was used for spectral background correction).

4.2.5 Total Free Thiol assay

The plasma sample was centrifuged for 10 minutes, and 75µL of the plasma extracted and mixed with a Tris pH 8.2 buffer. 90µL of prepared standard solution and plasma was added to the wells of the flat bottom well plate. The absorbance of this mix is measured at 412nm and 630nm, which is called absorbance pre-incubation. Then 20µL of DTNB (dithionitrobenzoic acid) was added to each well on the plate, put on the plate shaker and incubated at room temperature for 20 minutes, after which the absorbance of the samples is measured at 412nm and 630nm (absorbance post incubation). The assay was normalised by dividing by the protein content of each sample, measured using the Coomassie (Bradford) protein assay kit.

4.2.6 Statistical and regression analyses

The statistical testing and stepwise linear regression used in the analysis of assay data were performed using R (v.3.6.0) [178]. The Shapiro-Wilk test evaluates whether a random data sample forms a normal distribution. A significant result indicates a skewed distribution. The one-way analysis of variance (ANOVA) test indicates whether a significant difference exists in a continuous variable in two or more groups. The Kruskal-Wallis test is the non-parametric version of the one-way ANOVA. Regressions were performed with one or more variables. When more than one variable was included, the stepwise regression method was used. This combines the forward and backward selection of variables. When each new variable is added, all other current variables are examined. If any current variable is now non-significant, it is removed.

4.2.7 Supervised machine learning

For further analysis of antioxidant potential and reactive oxygen species assay results, supervised machine learning was used to classify extreme high and low assay results, using GenePy scores as prediction features. The genes chosen for inclusion as features in machine learning were based on NADPH oxidase gene literature [410]. The machine learning workflow is illustrated in Figure 25. The data set of each assay was separated into the top quartile ("high"), the bottom quartile ("low") and remaining results classified as "medium". Any results labelled as "medium" were removed from the data set. This created a balanced, binary classification problem, and it was expected that any machine learning algorithm would more easily distinguish between these two classes. This data set was split into training and testing data, in the ratio 80/20. Pre-processing of the data by centring and scaling (z-score) the GenePy scores was conducted on the training data and test data independently.

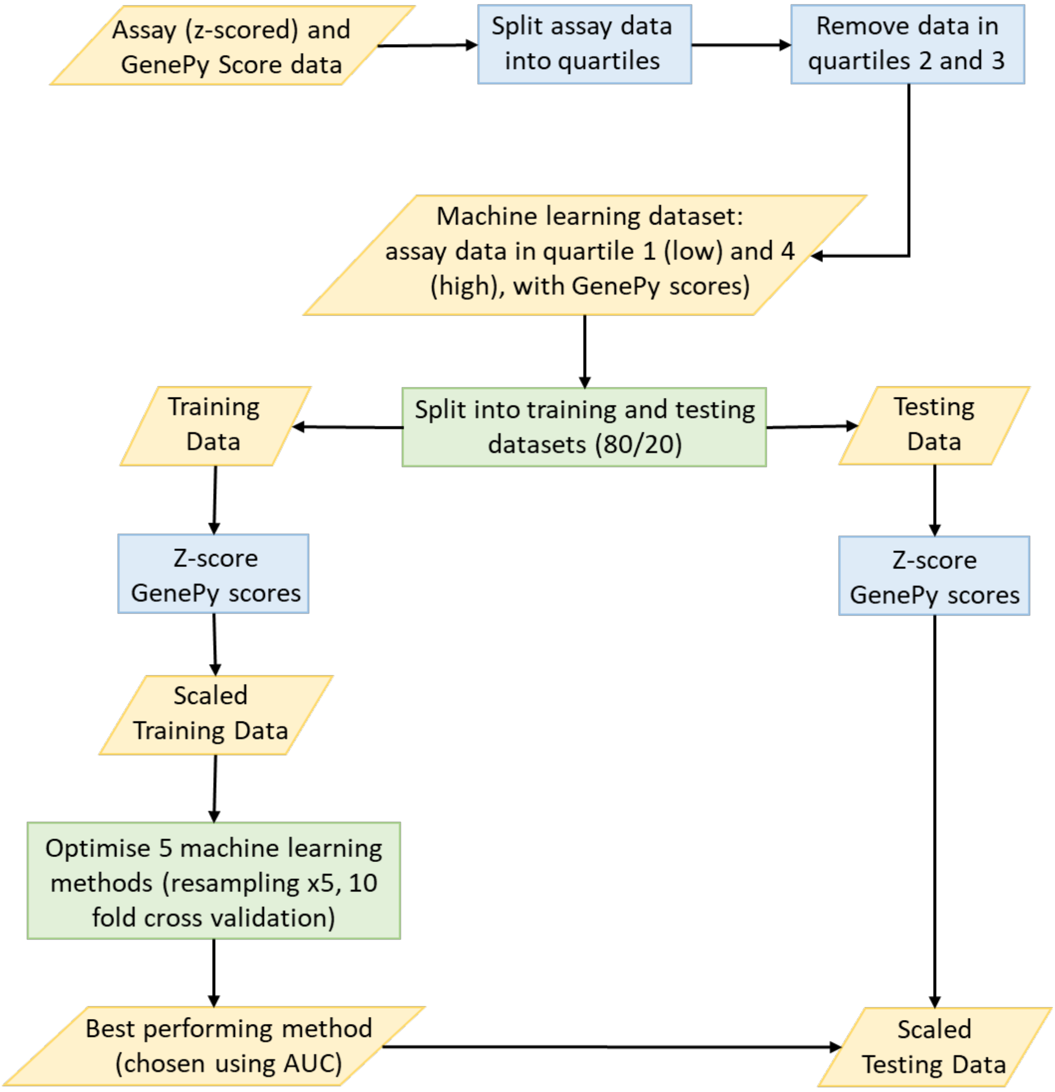

Assay (z-scored) and GenePy Score data
Split assay data into quartiles
Remove data in quartiles 2 and 3
Machine learning dataset: assay data in quartile 1 (low) and 4 (high), with GenePy scores)
Split into training and testing datasets (80/20)
Training Data
Testing Data
Z-score GenePy scores
Z-score GenePy scores
Scaled Training Data
Optimise 5 machine learning methods (resampling x5, 10 fold cross validation)
Best performing method (chosen using AUC)
Scaled Testing Data

Figure 25 Machine learning workflow for classifying extreme (high and low) assay data values using GenePy scores. The workflow was repeated for each assay.

All machine learning training and testing was accomplished using the caret package (v.6.0-84) [411]in R (v.3.6.0) [178]. Five machine learning algorithms were tuned, and their performance evaluated. These were two support vector machine algorithms, one with a linear kernel and one with a radial kernel, gradient boosting machines, random forest and logistic model trees. During training of each model, the data was resampled 5 times, and 10-fold cross validation was used.

The training process was repeated for each potential hyperparameter value in order to optimise each method. The support vector machine with radial kernel, and gradient boosting machines, had two tuning parameters, while the other three algorithms had one. The model with the highest AUC on the training data was chosen to classify the testing set. This was repeated for each assay independently (see Supplementary Files for machine learning scripts).

4.3 Results

The cohort for this study consisted of 331 patients, as this was the subset of patients from the Southampton Genetics of IBD study for which plasma samples were available for analysis (clinical characteristics available in Table 7). FRAP and TFT assay results are available for all probands, while the TBARS assay failed for four probands, and no duplicate is available for 25 probands. The TBARS assay failures are due to the miniaturisation of the assay, and the high temperature the samples were heated to, causing some samples to evaporate. The protein content measurement for the normalisation of the TFT assay was repeated for one proband due to a high variation between the triplicate samples (coefficient of variation 17.12%, median coefficient of variation over all samples 3.72%).

Table 7 Clinical characteristics of a paediatric cohort for which plasma samples were available, split by sex.

		Female	Male	Total
N		123	208	331
Median age at diagnosis (SD)		12 (3.25)	13 (3.44)	12 (3.36)
Range of age at diagnosis		2-16	1-17	1-17
Diagnosis	CD	73	155	228
	UC	42	41	83
	IBDU	3	6	9
	No IBD	5	6	11

To gain an initial understanding of these data produced by FRAP, TBARS and TFT assays, summary statistics are calculated (Table 8). The distribution of these data is visualised in Figure 26, and the

Shapiro-Wilk test was used to determine if the data of each assay deviates from a normal distribution. The TFT assay before and after normalisation have a normal distribution. However, there is significant evidence that the FRAP and TBARS assay results distributions deviate from normality (p=1.215e-14, p=4.824e-11, respectively). Therefore, non-parametric statistical tests were required for subsequent analyses of these data.

Table 8 Basic statistics describing results from TFT, FRAP and TBARS assays.

	FRAP	**TBARS**	**TFT**	**TFT (Normalised)**
N	331	327	331	331
Mean	1012.039	7.325	375.389	6.147
Median	961.000	6.910	375.268	6.130
Standard Deviation	262.189	2.919	71.537	0.998
Range	1949.000	22.600	441.372	5.940
Minimum	506.000	1.600	191.764	3.420
Maximum	2455.000	24.200	633.136	9.360

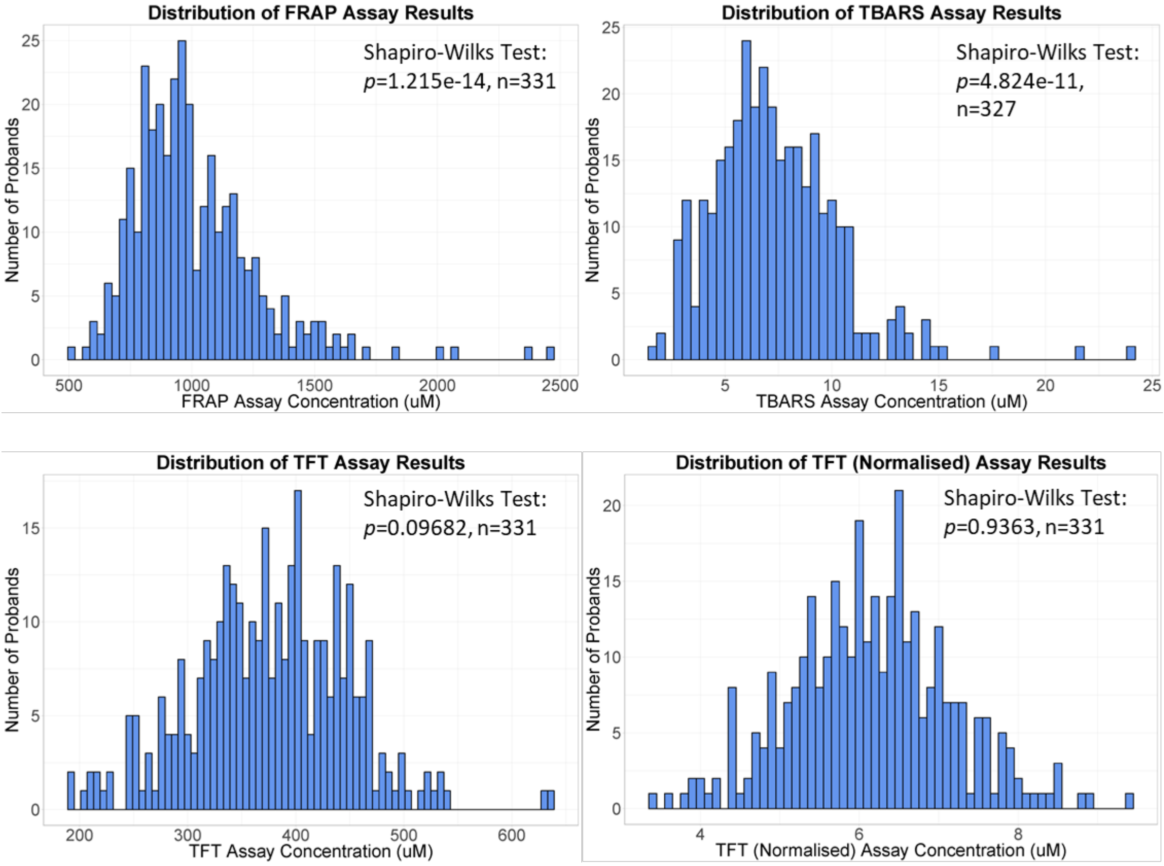


Figure 26 FRAP, TBARS, TFT before normalisation and TFT after normalisation assay results distribution. FRAP and TBARS assay results have a (positively) skewed distribution, confirmed by the Shapiro-Wilks test for normality. TFT assay results before and after normalisation has a normal distribution.

4.3.1 Plate analysis and correction

The distribution of the assay data was tested to determine if assay results were in any way influenced by batch effects. Kruskal Wallis test for the FRAP and TBARS assays and one-way ANOVA test for the TFT assay were used to assess differences between result distributions across plates, and across days. The associated p-values indicated significant differences in plates for all assays, and significant differences in days for TFT and FRAP (Figure 27), indicating that some correction needed to be applied to the data to prevent bias in subsequent results due to batch effects. The FRAP assay plate in particular displayed a strong trend of the median result on a plate increasing in each subsequent plate. This trend is present for all three days the FRAP assay was conducted. The results from testing the distribution on the TBARS plates and days informed how the correction should be implemented. In this case the differences between plates were masked when the data for the overall day was observed. This demonstrated that any correction must be done in reference to plates, rather than days.

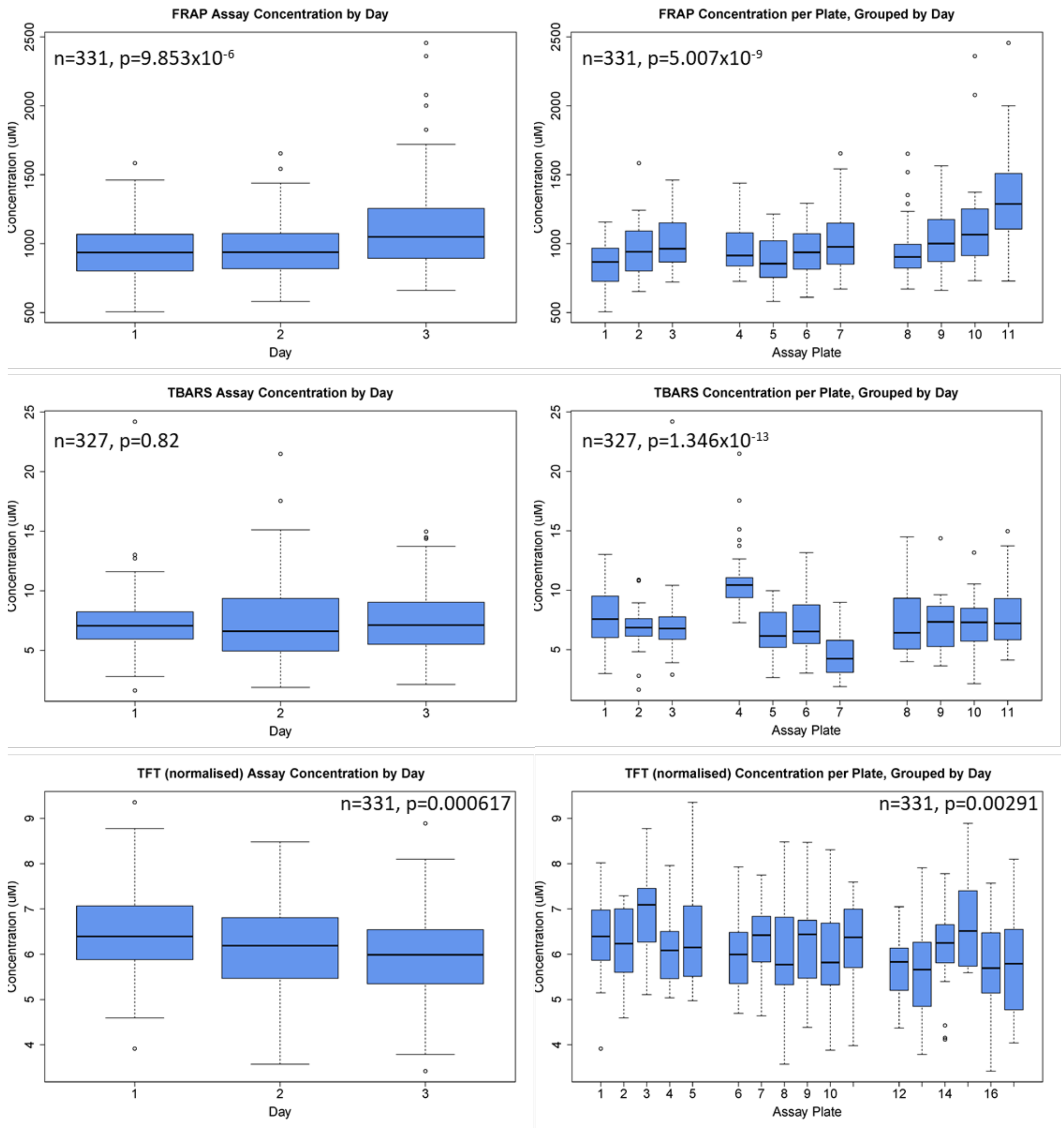


Figure 27 FRAP, TBARS and TFT assay concentrations by day (left) and by plate (right). Kruskal-Wallis test (FRAP, TBARS) and one-way ANOVA (TFT) p-values noted on each plot for differences in assay concentrations per day and per plate.

As oxidative stress can increase as the body ages, a potential cause of significant batch effects would be that some plates were enriched for patients of younger or older age. Kruskal-Wallis and one-way ANOVA tests revealed no significant differences in the age distribution on each plate (Table 9). Given the cause of significant batch effects could not be determined, the raw assay results were transformed into z-scores within batches. This conserved the extreme values observed within each assay (Figure 28).

Table 9 Results for determining whether a significant difference exists in the ages of the samples on each plate.

	N	Test	*p* value
TFT plate vs Age at blood draw	331	One-way ANOVA	0.200
FRAP plate vs Age at blood draw	331	Kruskal-Wallis	0.798
TBARS plate vs Age at blood draw	327	Kruskal-Wallis	0.775

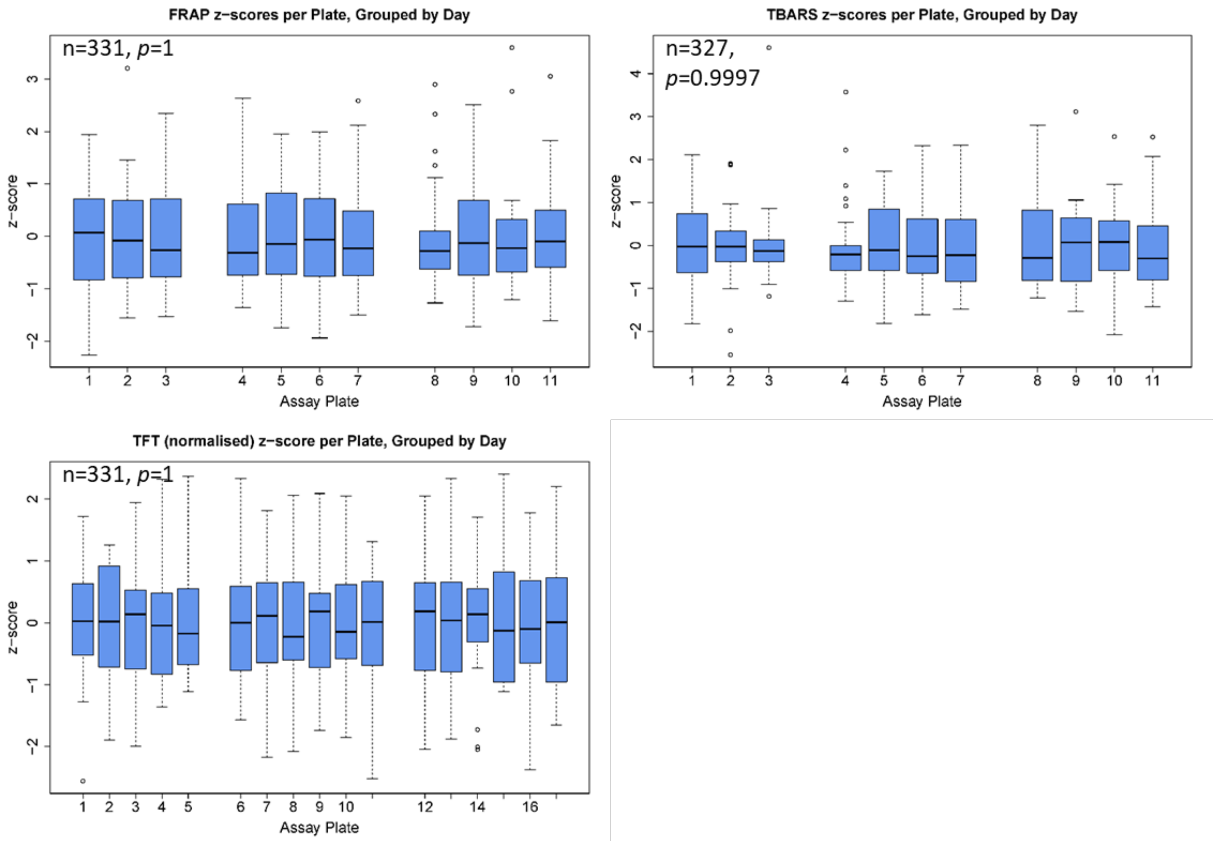


Figure 28 Boxplots of FRAP, TBARS and TBARS assays after z-score conversions. Kruskal-Wallis test (FRAP, TBARS) and one-way ANOVA (TFT) p-values included.

4.3.2 Correlation between oxidative stress assays

It was hypothesised that patients with lower z-scores in the FRAP (antioxidant potential) assay would be ill-equipped to effectively deal with increased oxidative stress, resulting in higher z-scores in the TBARS and TFT assay. A linear regression was performed between TBARS and FRAP assay results and TFT and FRAP assay results (Figure 29). A significant linear correlation assists between FRAP and TFT assays, although it accounts for very little of the variation in the data (adjusted R^2=0.0284). The direction of this correlation does not support the hypothesis. The same analysis conducted only on those cases with CD (n=225) results in a slight increase in R^2 for the significant regression of TFT versus FRAP (adjusted R^2=0.0295, p=0.005912). Additionally, when this regression is performed on only UC data (n=81), this relationship does not persist. This is either due to fewer cases of UC in the dataset, or because the biological connection between reactive oxygen species and UC is weaker, or not at all present.

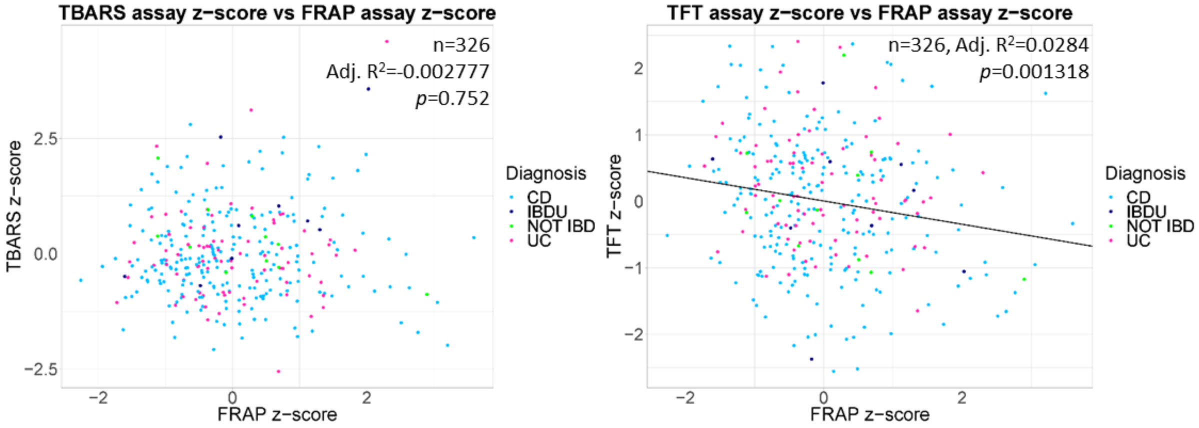


Figure 29 TBARS assay z-score vs FRAP assay z-score, and TFT assay z-score vs FRAP assay z-score, with details of the linear regressions in the top-right of each figure. The regression line is shown where the adjusted R^2 generated from the linear regression is significant.

4.3.3 Oxidative stress assay results and inflammatory bowel disease diagnoses

The oxidative stress assay profile across patients with different diagnoses of IBD was examined (Figure 30). No significant differences between the assay results of different diagnostic groups were found using the Kruskal-Wallis (FRAP, TBARS) and one-way ANOVA (TFT) statistical tests.

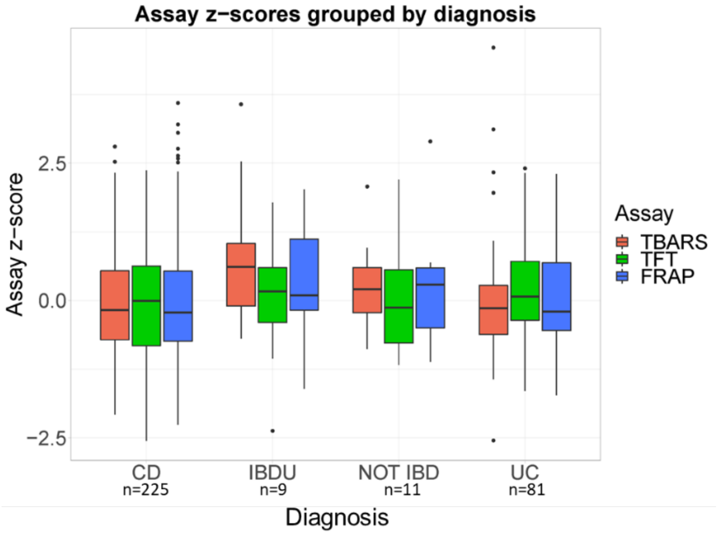


Figure 30 FRAP, TBARS and TFT assay distributions, grouped by diagnosis.

4.3.4 Assays and age

To assess whether a linear relationship exists between assays and age of diagnosis, and between assays and age at blood draw, linear regressions between these variables was calculated. These results are shown in Figure 31, which also allows the comparison of the distributions of the data. When this analysis is conducted on the CD data (n=224), the same adjusted R^2 are obtained. This suggests this relationship between age and assay result is driven by the CD cases. To determine whether age at diagnosis was significantly associated with assay results, or simply related to patient age at blood draw, a stepwise linear regression was used. It was thought that by using this method the more significant age variable could be determined. The results of this are shown in Table 10. The TFT assay has no significant correlation with either age variable. FRAP and TBARS assays both have significant relationships with age variables. The regression for the TBARS assay indicates that both age at blood draw and age at diagnosis are significant, although the corresponding adjusted R^2 for this regression is very small. Therefore, from these results it is more likely the age at blood draw is the influential variable.

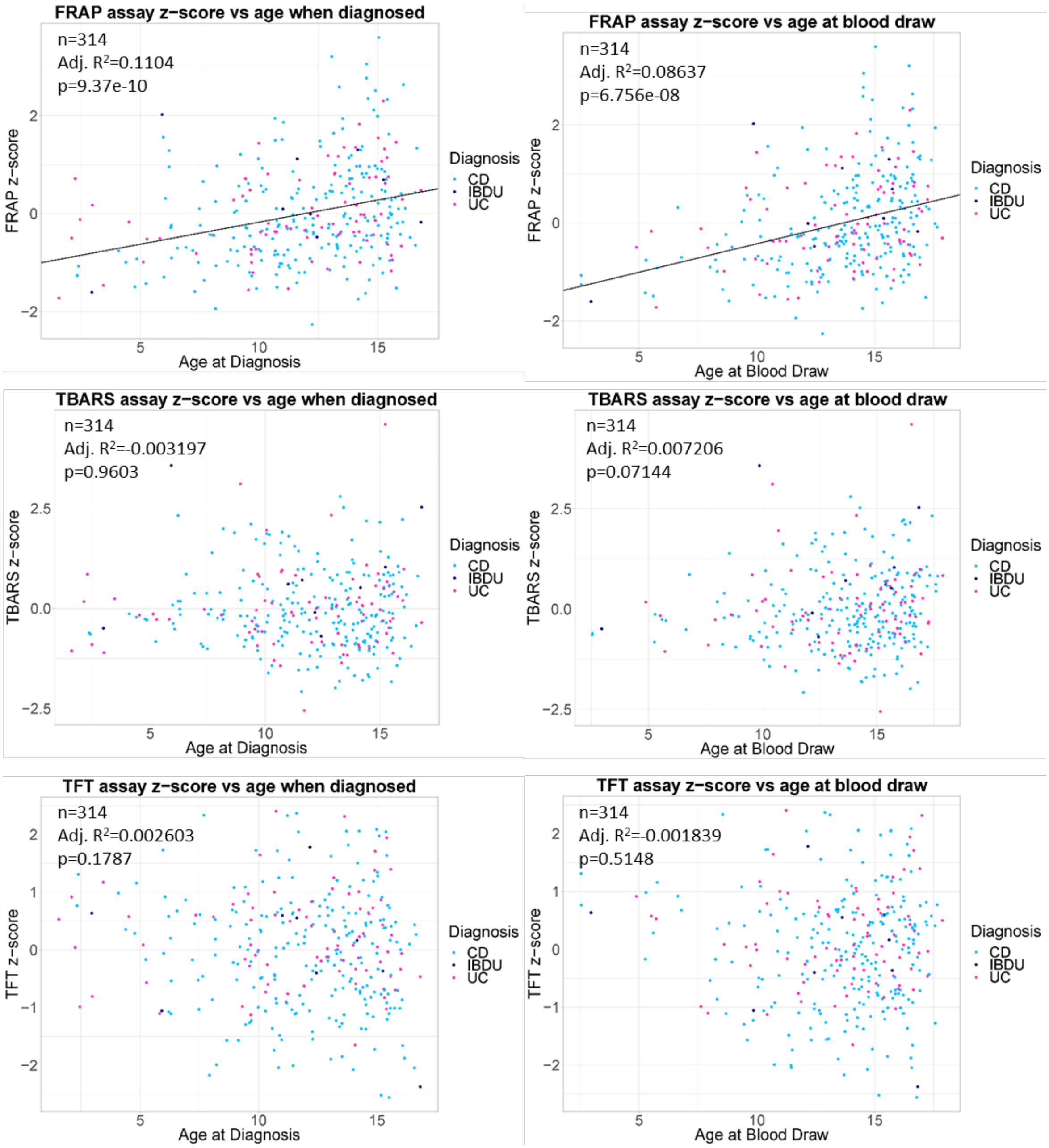


Figure 31 Plots of assays versus age at diagnosis, and age at blood draw. Details of the results of the linear regression are displayed in the top-left corner. Where the R^2 generated was significant, the corresponding regression line is shown in the figure.

Table 10 Results of the stepwise linear regression for assay and age variables.

Assay	Variable	Coefficient (*p* value)	Adjusted R^2 (*p* value)
FRAP	Age at blood draw	0.09306 (***0.00244***)	0.1101 (***4.862x10^{-9}***)
	Age at diagnosis	0.02519 (0.34514)	
TBARS	Age at blood draw	0.09872 (***0.00219***)	0.02354 (***0.009085***)
	Age at diagnosis	-0.06971 (***0.01316***)	
TFT	Age at blood draw	0.02239 (0.486)	0.0009588 (0.3179)
	Age at diagnosis	-0.03844 (0.172)	

4.3.5 Oxidative stress assays and C-reactive protein

C-reactive protein (CRP) is a commonly tested marker of inflammation, which can be used to monitor disease [406]. The majority of available data on CRP concentration was from probands with CD (CD=145, UC=11, IBDU=2). To compare inflammation and oxidative stress, a linear regression model was used where each assay's z-score was regressed against the CRP results (Figure 32). It was expected that patients with a high CRP result would also have high assay results. A significant linear relationship with CRP exists for those assays that measure oxidative stress, TFT and TBARS. This relationship was not what was expected. Instead, those patients with high CRP concentrations have lower results in the oxidative stress assays.

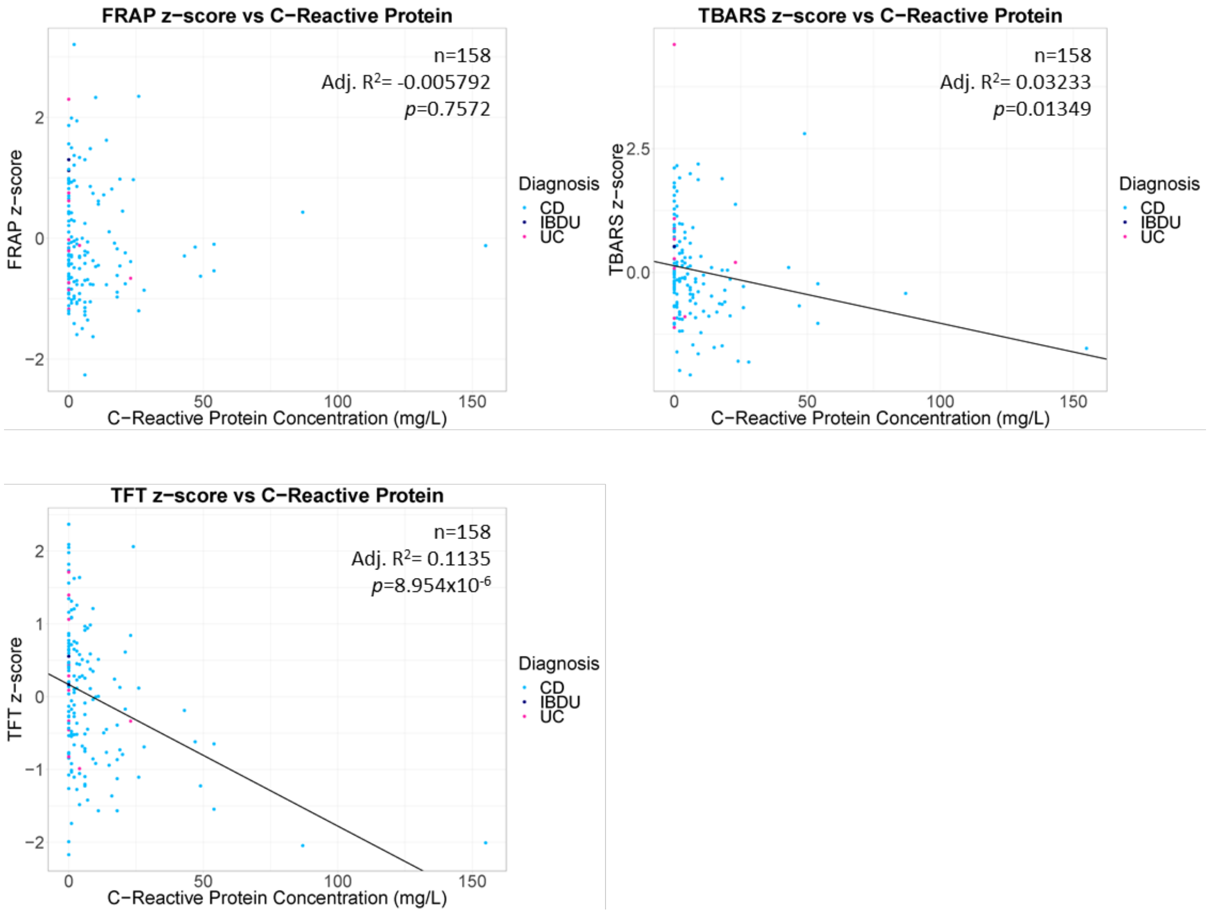


Figure 32 FRAP, TBARS and TFT assay z-scores plotted against CRP concentration. Where the p value of the adjusted R^2 was significant, the regression line representing the linear model is shown.

4.3.6 GenePy scores

GenePy scores were generated for 15 NADPH oxidase genes, and 3 genes of interest thought to have a role in oxidative stress or inflammation. For one potentially important NADPH oxidase gene, *NCF1*, as there was no coverage of that gene in version 4 of the sequencing library. As 83 probands were sequenced with this version, a large proportion of patients would not be included in the analysis if the GenePy scores were generated using only the data from later versions of the sequencing library. Therefore, this gene was omitted. Although GenePy scores were generated, there were no exonic variants in the gene *RAC1*, resulting in all zero GenePy scores, as these were based on exonic variants present in the gene. Coverage and function of each gene and potential relationships to IBD are contained in Table 11.

Table 11 Coverage of genes used in analysis (Agilent SureSelect Human All Exon v 4, 5 and 6). Information on gene function and any literature evidence of contributions to IBD development caused by these genes. Tier 1 genes have variants that have been implicated in IBD, tier 2 genes have not been implicated in IBD, but are NADPH oxidase genes. Some tier 2 genes code for proteins that form complexes with proteins encoded by genes in tier 1.

Gene	**Coverage**			**Gene Function**	**Relation to IBD**
	V4	V5	V6		
Tier 1					
NOX1	0.615	0.620	0.528	The complex is responsible for one-electron transfer of oxygen to generate superoxide.	Loss-of-function variants in NOX1 can be context-specific disease modifiers [412].
CYBA	0.311	0.283	0.256	Part of the NOX1, NOX2 and NOX3 enzyme complexes	These genes are implicated in CD. Linked to a higher likelihood of perianal disease and stricturing disease [110].
CYBB/NOX2	0.621	0.643	0.538	Part of the NOX4 enzyme complex	
NCF1	0	0.499	0.455	Forms NOX2 enzyme complex, with NCF2, NCF4, RAC2 and RAP1A.	
NCF2	0.498	0.511	0.445	Forms NOX2 enzyme complex, with NCF1, NCF4, RAC2 and RAP1A.	
NCF4	0.626	0.565	0.483	Forms NOX2 enzyme complex, with NCF1, NCF2, RAC2 and RAP1A.	

DUOX2	0.563	0.613	0.549	Protein encoded forms similar enzyme complex to the NADPH oxidase complexes, but the end product is hydrogen peroxide.	Missense variants identified in VEOIBD patients, showing reduced ROS production [108]
REG3A	0.702	0.661	0.548	Associated with cell proliferation or differentiation, anti-inflammatory. Part of the REG gene family (REG1, REG2A, REG2B, REG3A, REG4). Mediates killing of gram-positive bacteria. Regulates keratinocyte proliferation and differentiation after skin injury.	Increased expression in IBD [413]. In-house machine learning analysis suggested REG3A as a gene of interest
HMOX1	0.737	0.697	0.647	HMOX1 is a component of antioxidant defence against oxidative stress [414].	In-house variant analysis implicated this gene in CD
NOD2	0.720	0.664	0.609	Recognises bacterial lipopolysaccharides, activates NF-κB and IFN-β pathways [102].	A monogenic CD gene, and also causes CD in combination with other genes [5, 415].
Tier 2					
NOXA1	0.514	0.546	0.475	Activation of NADPH oxidases	
NOXO1	0.748	0.671	0.674	Activation of NADPH oxidases	
RAC1	0.543	0.557	0.541	Present in NOX1 enzyme complex	

RAC2	0.445	0.419	0.373	Present in NOX2 enzyme complex	Potential association with CD [416].
DUOXA1	0.808	0.781	0.636	Gene encodes maturation factor for DUOX1 function	
DUOXA2	0.708	0.718	0.659	Gene encodes maturation factor for DUOX2 function	
DUOX1	0.524	0.596	0.469	Similar enzyme complex to DUOX2, found in different tissues/cells.	
NOX3	0.622	0.683	0.535	Similar enzyme complex to NOX1/NOX2, found in different tissues/cells.	
NOX4	0.407	0.503	0.444	Similar enzyme complex to NOX1/NOX2 found in different tissues/cells.	
NOX5	0.693	0.720	0.593	Similar enzyme complex to NOX1/NOX2, found in different tissues/cells.	

4.3.7 Stepwise linear regression

A stepwise linear regression was used to determine to what extent a linear relationship existed between antioxidant potential and oxidative stress assay results and genes identified as potentially related to reactive oxygen species production and oxidative stress. Six models were created, two per assay, with one using all available data, and another only using data from probands diagnosed with CD, following on from previous results indicating these relationships may be more important in these patients. Seventeen GenePy scores were used as predictor variables. The stepwise regression used both forwards and backward feature selection, to obtain the best combination of GenePy scores for assay result prediction (Table 12).

Table 12 Stepwise linear regression results for FRAP, TBARS and TFT, for two datasets.

Assay	Data (n)	Variable	Coefficient	*p* value	Adjusted R^2 (*p* value)	*p* value
FRAP	All Data (314)	*CYBA*	-0.541	0.094	0.0099	0.078
		NOX5	0.418	0.142		
	CD data only (224)	*DUOXA1*	-0.725	0.148	0.0088	0.139
		NOX5	0.577	0.118		
TBARS	All data (314)	*RAC2*	1.022	***0.013***	0.03296	***0.0062***
		REG3A	1.457	0.152		
		DUOXA1	-0.892	***0.045***		
		DUOXA2	1.026	0.122		
	CD data only (224)	*REG3A*	1.468	0.136	0.03537	***0.012***
		DUOXA1	-0.939	***0.040***		
		DUOXA2	1.606	***0.021***		
TFT	All data (314)	*HMOX1*	-0.674	0.052	0.07958	***5.65×10^{-6}***
		NOXA1	-0.905	***0.001***		
		DUOX1	-1.138	***1.07×10^{-4}***		

		NOX5	0.430	0.123		
	CD data only (224)	*HMOX1*	-0.915	***0.018***	0.08161	***7.33x10^{-5}***
		NOXA1	-0.942	***0.005***		
		DUOX1	-0.914	***0.007***		

There is no significant adjusted R^2 for the regression with either FRAP assay dataset, but there are significant regressions for TBARS and TFT assay data (TBARS adjusted R^2 0.03296 (all data), 0.03537 (CD data only), TFT adjusted R^2 0.07958 (all data), 0.08161 (CD data only)). Where the adjusted R^2 was significant, an increase in this statistic is observed when only CD data is used. These linear relationships do not account for much of the variation in the data, indicating that more complex methods may produce better predictions.

4.3.8 Machine learning for predicting assay results

All identified NADPH oxidase genes, where GenePy scores were available and not invariant were included as features (n=15): *NOX1*, *CYBB*, *CYBA*, *NCF2*, *NCF4*, *RAC2*, *DUOX2*, *NOXA1*, *NOXO1*, *DUOXA1*, *DUOXA2*, *DUOX1*, *NOX3*, *NOX4*, and *NOX5*. The five machine learning algorithms: gradient boosting machines (GBM), logistic model trees (LMT), random forest (RF) and support vector machines with radial and linear kernels (SVM (R) and SVM (L), respectively) were run separately for each assay's training data (n=128 per assay dataset). Training involved optimising each method on the training data. Details of the hyperparameters, the values of the hyperparameters trialled, and the optimum parameter values for each assay are given in Table 13. The AUC values produced during training are visualised in Figure 33A-C. With FRAP and TBARS assay training data, SVM (R) was the best performing model with the highest AUC. For the TFT assay training data, the GBM model had the highest AUC (Table 14).

Table 13 Description of the tuning parameters for each model and optimal parameter values in training for the prediction of the extreme assay results in each assay's data set.

Model	**Hyperparameter**	**Hyperparameter Description**	**Hyperparameter Values Tried**	**Optimal Parameter Value**		
				FRAP data	TBARS data	TFT data
GBM	interaction.depth	Maximum tree depth	1,3,5,7,9	5	3	9
	n.trees	Number of boosting iterations	50-1,500 in intervals of 50	400	1200	50
LMT	iter	Number of iterations	1, 20, 40, 60, 80, 100, 150, 200, 250, 300, 400, 500	100	400	60
RF	mtry	Number of randomly selected features	2, 8, 15	15	2	2
SVM (L)	C	Cost	0.75, 0.9, 1, 1.1, 1.25, 1.5, 1.75	0.9	1.1	1.25
SVM (R)	sigma	Sigma	0.01, 0.015, 0.2, 0.25	0.25	0.25	0.2
	C	Cost	0.75, 0.9, 1, 1.1, 1.25	0.75	0.9	0.75

Table 14 AUC achieved by each model for each assay training dataset.

	AUC, Training Data		
	FRAP assay	TBARS assay	TFT assay
GBM	0.4659410	0.5352494	***0.5730952***
LMT	0.4889116	0.5601984	0.5390306
RF	0.3563039	0.5644501	0.5517857
SVM (R)	***0.6224830***	***0.6075624***	0.5314399
SVM (L)	0.4661565	0.5435601	0.5121995

These machine learning models, each with hyperparameters selected for to maximise AUC, were used on the testing data (n=32 for FRAP and TFT assay data, n=30 for TBARS assay data). When applied to the test data sets, the models for TBARS and TFT failed to predict high and low assay results (TBARS accuracy=0.33, sensitivity=0.40, specificity=0.27; TFT accuracy=0.41, sensitivity=0.40, specificity=0.27). The SVM (R) for the FRAP assay data was a reasonable predictor and was better at identifying the assay results in the upper quartile (sensitivity=0.69, specificity=0.56). However, there were very wide confidence intervals for the accuracy of this model: accuracy=0.63, 95% CI 0.44 – 0.79. AUCs for each assay's testing data are visualised in Figure 33D-F.

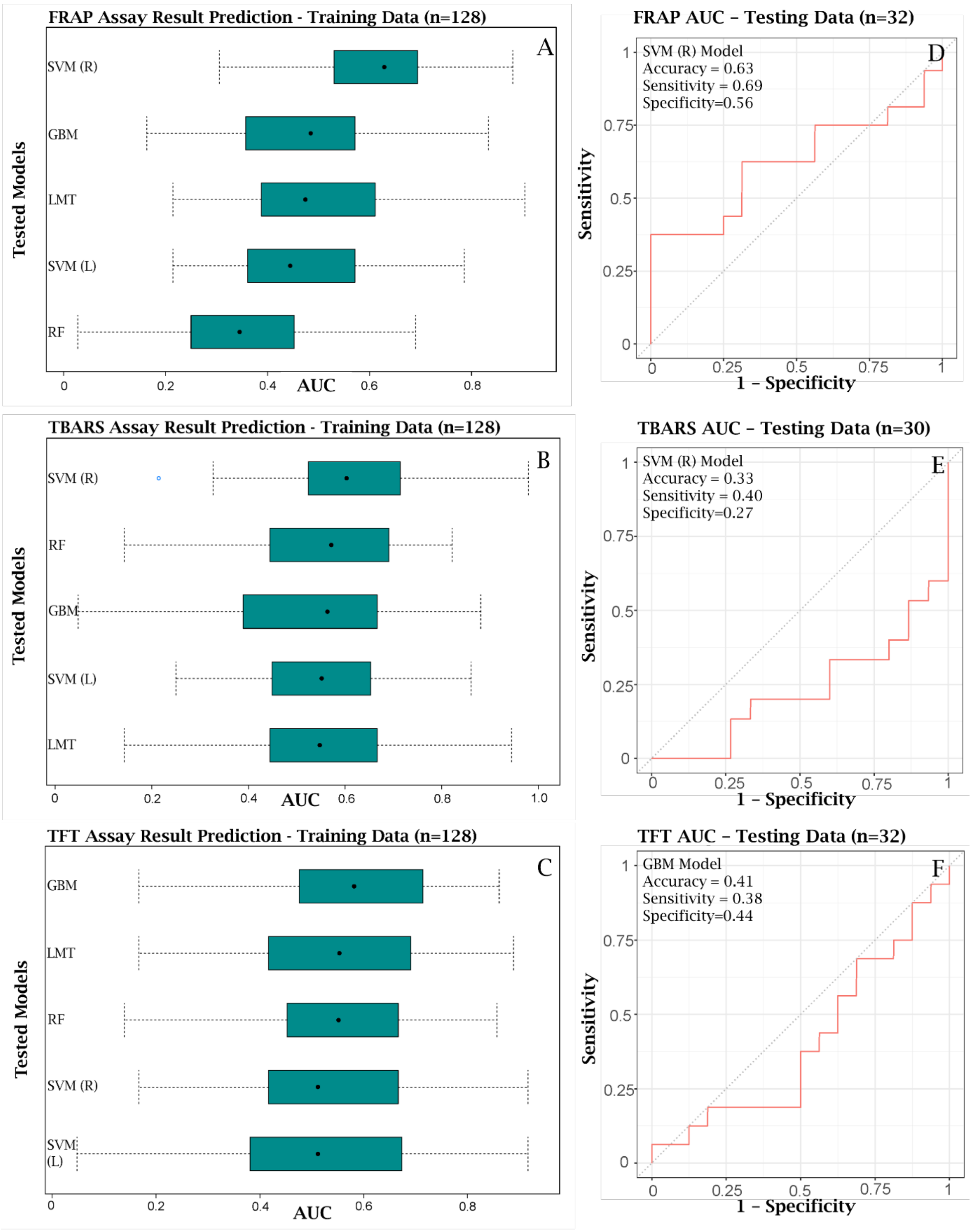


Figure 33 Training and testing machine learning models for differentiating patients with extreme assay results using GenePy scores. A-C: Five models trained on a balanced two class datasets (n=128) for each assay: SVM (R), SVM (L), GBM, RF and LMT. Boxplots sorted by performance (top to bottom). D-F: Best performing model on the training data applied to the test data, prediction performance demonstrated with AUCs.

4.4 Discussion

Here, data from the oxidative stress assays TFT and TBARS, and the antioxidant potential assay FRAP were explored to first understand these data, then understand potential relationships between assays and IBD, and finally the relationships between assays and genetic variation (using GenePy scores). Linear regressions between FRAP and TBARS, and FRAP and TFT assays revealed little overlap between oxidative stress and antioxidant potential assays, illustrated by low adjusted R^2 values (although this was significant for the FRAP and TBARS regression). There was a negative correlation between TFT and FRAP that was unexpected. In a "healthy" control population this negative relationship would be expected, an increased antioxidant potential leading to lower general oxidative stress. Therefore, it was expected that in the IBD cohort there may be some perturbation from this. As there was no control cohort, it may be that the relationship observed is milder than that present in a control cohort, indicating some dysregulation.

There were statistically significant differences observed between the assay results on different plates for the FRAP and TBARS assays. This was corrected by transforming results into z-scores within each plate. For the FRAP assay in particular there appeared to be an upward trend in results for each plate, and these upwards trends were observed for each day, with assay results appearing to return to lower values at the start of each day. After consultation with the team that performed the lab work, the time of day was thought to be a factor, potentially combined with temperature. One person performed the FRAP assay for each plate sequentially over the 3 days, and temperature may have been a factor as the FRAP assay protocol states that plasma samples were thawed when necessary. Samples thawed later in the day may have been brought up to temperature more quickly due to a higher ambient temperature. For the TBARS assay, there is less of a clear trend. The significant differences between plates are no longer significant after observing results obtained per day. However, this assay was miniaturised, which led to a protocol that was more difficult to execute, which may account for inconsistencies in results.

No statistically significant differences were observed when assay results were compared across diagnosis groups, However, there were outliers in the data, particularly in the TBARS assay data for CD and UC, and in the FRAP assay data for CD. It is not expected that all patients within either disease subtype have the same aetiology. There may be a small subset of patients that have a distinctive ROS assay profile that is indicative of oxidative stress being the primary contributor to disease aetiology. Further investigation of outliers may consist of an analysis of genetic variation that could underpin these comparatively extreme assay results.

The investigation of associations between the FRAP assay and age revealed that these observed differences were likely to be driven by the patient's age when the blood was taken. This is as opposed to the age at diagnosis, as the plasma for assays were not necessarily taken on or near the day of diagnosis. Although there is limited evidence, one paper suggests that an increase in antioxidant potential may be standard in a paediatric cohort. In an analysis of antioxidant capacity in children with childhood caries, the whole cohort was used to plot the total antioxidant capacity in saliva against age [417]. This produced a positive linear regression that was close to significant, similar to the significant one viewed in this analysis. This regression could have perhaps been significant with a larger cohort, or a wider age range, as 100 children were included between 3 and 5 years of age. This leads to a tentative suggestion that antioxidant potential may naturally increase as children age, and their bodies become more capable of handling oxidative stress. When considering the TBARS assay and age, the correlation between age at blood draw and lipid peroxidation was negative, while the correlation between age at diagnosis and lipid peroxidation was negative. While these correlations, and the adjusted R^2 were significant, the contradictory directions of these correlations mean the evidence is not conclusive, especially when also considering the other measure of oxidative stress (TFT) had no significant correlations with age. For all assays, the results indicate that these measures cannot be used as an indicator for early onset IBD.

No significant correlation between CRP and FRAP was observed. There were small but significant correlations for the TBARS and TFT assays. In these cases, an increased CRP was associated with decreased oxidative stress assay results. This was contrary to expectations, as it was thought that an increase in reactive oxygen species production, leading to oxidative stress would result in higher inflammation. The results here appear more in line with the pathology of Chronic Granulomatous Disease, where loss of function variants cause a decreased production of reactive oxygen species [104].

There was very limited evidence that accrued pathogenic variation in genes that might impact proteins in oxidative stress pathways, were correlated with oxidative stress assays. The highest correlation (R^2=0.082) was between GenePy scores and TFT assay on the CD patient subset. The GenePy scores for the genes *HMOX1*, *NOXA1* and *DUOX1* were the factors identified that explained some of the variance in the TFT assay data. *HMOX1* is a component of antioxidant defence against oxidative stress [414]. *NOXA1* is an activator of *NOX1*, and *DUOX1* is a component of an NADPH oxidase complex that is known to be expressed in the colon [105]. Instead of the genetic burden existing in one gene, variation in a combination of genes involved across this pathway could be contributing to the function of downstream complexes, reflected in the oxidative stress assays. Sensitivity to detect a relationship between mutations in genes encoding proteins critical to ROS and anti-oxidant potential may be improved by also including the sum of GenePy scores that are part of one complex. For example summing all scores of genes that code for proteins in the *NOX1* complex. This has been shown as a promising approach in analyses of GenePy scores and transcriptomic data (unpublished data). This method takes into account potential interactions between complexes that may, in combination, cause an effect such as dysregulation of reactive oxygen species production. There was no significant relationship between GenePy scores and the FRAP assay in the regression analysis. This is in contrast to the machine learning results, where the FRAP assay classifier performed the best. This could be indicative of nonlinear relationships between GenePy scores for all assays that cannot be identified through regression analysis.

Machine learning was applied to attempt to differentiate between high and low assay results using GenePy scores. The prediction problem was adjusted from a regression to a two-class classification, and results in quartiles one and four of each assay distribution were retained to make these categories (high and low) more distinct. Despite this, the machine learning models for the TBARS and TFT assays performed poorly, and the model for FRAP assay results was only modestly good. Although the genes used as features were selected using biological knowledge, the feature set was small for each of the three models (n=15). Additionally, not all features had been implicated in IBD. The machine learning may have benefitted from an initial feature selection step, and maybe expanding the genes included to other, related pathways upstream or downstream of ROS production. However, it is not unreasonable to suggest that the ability for machine learning to predict assay results from the GenePy scores is weak. Many factors may affect the connection between the features and the outcome variable, including transcriptomic, diet and environmental factors. The machine learning classifier for the FRAP assay did indicate that these results are more strongly derived from genetics. Future work here may include incorporating these assay results as a feature, along with other clinical features such as blood test results in a potential model for CD. This could address disease activity monitoring, or complications of CD. The current volume of assay data for CD is however quite small (n=224), and may not be adequate to train and test a machine learning model once the data is split according to the particular outcome variable.

There are a few limitations in the assay data used here. The first is that the data did have to be transformed into z-scores due to inconsistencies across the different plates. Although z-scores preserve the extreme results within the new range, it is possible that potentially useful information was lost during this process. Secondly, it is not known how much the results of these assays for each patient would vary over time, and what could cause this variation. The individual's diet and environment could affect these results, as could the treatment they are on, and the activity of disease. The data are a cross-sectional snapshot, and not longitudinal. Finally, these data were generated from blood plasma. It is not clear whether an under or overproduction of reactive oxygen species in the gut would be reflected in the plasma.

Although the analyses in this chapter have not revealed any assay to be a biomarker for IBD, it adds further evidence to current biological knowledge. A consistent theme across the analyses was that the results showed a stronger tendency towards significance when the analysis was restricted to Crohn's Disease patients. This was demonstrated in linear regressions between assays and age, and when stepwise linear regression was used for predicting assay results from GenePy scores. These findings are in line with Jahanshahi et al.'s results using the TBARS assay [405], but corroboration with other literature is limited by research often focusing on cases versus healthy controls. These relationships are likely only prevalent in a subgroup of CD patients where dysfunction in reactive oxygen species activation pathway contributes to pathology. Further stratification of the CD subgroup may be necessary.

As was described in Section 1.2.6.3, pathways involving *NOD2* signalling are somewhat interlinked with reactive oxygen species production by NADPH oxidase complexes. Variants in *NOX1* and *CYBA* were identified in a patient with VEOIBD, and functional assay analysis confirmed the protein encoded by *CYBA* (p22phox) did interact with *NOD2* [418]. Mouse models have also identified that deficiency of *NOD2* together with *CYBB* causes intestinal inflammation like the type and pattern of CD [419]. Evidence of interactions between *NOD2* and NADPH oxidase genes, combined with the observed relationships between the assay results and CD further evidences that reactive oxygen species and NADPH oxidase analysis will be most profitable using a CD subtype cohort.

Chapter 5 Random forest classification of inflammatory bowel disease subtypes and the Crohn’s disease stricturing endotype

Chapter summary – this chapter is the first of three chapters dedicated to stratifying IBD patients into clinical groups, using genomic data and ML. Initial random forest model results are obtained for two clinical tasks: classifying IBD patients into CD and UC, and classifying CD patients into stricturing and not-stricturing groups. These classification tasks are performed using three different gene panels. Additionally, two different forms of the GenePy matrix were investigated, along with an additional pre-processing step, to see if the performance of the random forest could be improved with changes to the input genomic data.

Chapter contributions – Whole exome sequencing data was joint called by Guo Cheng, with all subsequent processing, and transformation into GenePy scores, performed by Imogen Stafford. The IBD gene panel was curated by Guo Cheng and James Ashton. Extraction of clinical data from University Hospital Southampton records was performed by Florina Borca and Hang Phan. Clinical stricturing status was assessed by Melina Kellerman, Imogen Stafford and James Ashton. Fuentes false positive gene list remapping was performed by Ellie Seaby and Imogen Stafford. Ellie Seaby also assisted with quality control checks. ML pipeline was generated by Enrico Mossotto and Imogen Stafford.

Supplementary files can be found at https://doi.org/10.5258/SOTON/D2655.

5.1 Introduction

Early diagnosis is important for many chronic diseases, including inflammatory bowel disease. Diagnosing individuals with the correct subtype is crucial, in order that the patient receives treatment for the induction and maintenance of remission specific to that subtype. Furthermore, a delayed subtype diagnosis can result in an increased risk of complications that can require surgery [420, 421]. In paediatric cases, a delay of over 8.8 months was shown to be independently associated with impaired growth that persisted one year after diagnosis [21]. In particular, studies have found that it takes longer to obtain a diagnosis of CD for both paediatric and adult cases [21, 420]. The median diagnosis time in a paediatric cohort was 2.4 months for UC/IBDU (combined in study analysis) and 6.8 months for CD [21], and there were similar median diagnostic times in an adult cohort (CD 5 months, UC 1 month) [420]. It has been hypothesised this is due to many CD symptoms overlapping with other diseases, whereas specific UC symptoms, such as bloody diarrhoea, have been shown to decrease the likelihood of diagnostic delay [21]. In addition, some delays in diagnosis are caused by a lag between primary care referrals and an appointment with a specialist clinician [422]. One cohort study also found that a previous diagnosis of Inflammatory Bowel Syndrome or depression resulted in an increased wait for referral to a specialist [422]. In the UK, the National Institute for Health and Care Excellence states that no patient should wait more than four weeks to be seen by a specialist [423]. Aside from this, there is very limited clinical guidance regarding timelines for diagnosis. There are no recommendations for time to complete each diagnostic assessment by in the revised Porto criteria for paediatric patients [7]. In the British Society of Gastroenterology consensus guidelines, which govern adult IBD diagnosis and management [8], there is only a recommendation that a full ileocolonoscopy be conducted within the first year, to definitively confirm subtype diagnosis, and assess disease extent and severity. Some clinical investigations are conducted in order to eliminate other diseases a patient may have, for example coeliac disease testing, primary sclerosing cholangitis, and functional gut disorders [424, 425]. Details of common clinical investigations to diagnose IBD, and more specifically CD and UC, are given in Figure 34. In cases where symptoms are general like CD, using different patient information such as genomic and immunologic data to diagnose may be beneficial for reducing diagnosis times.

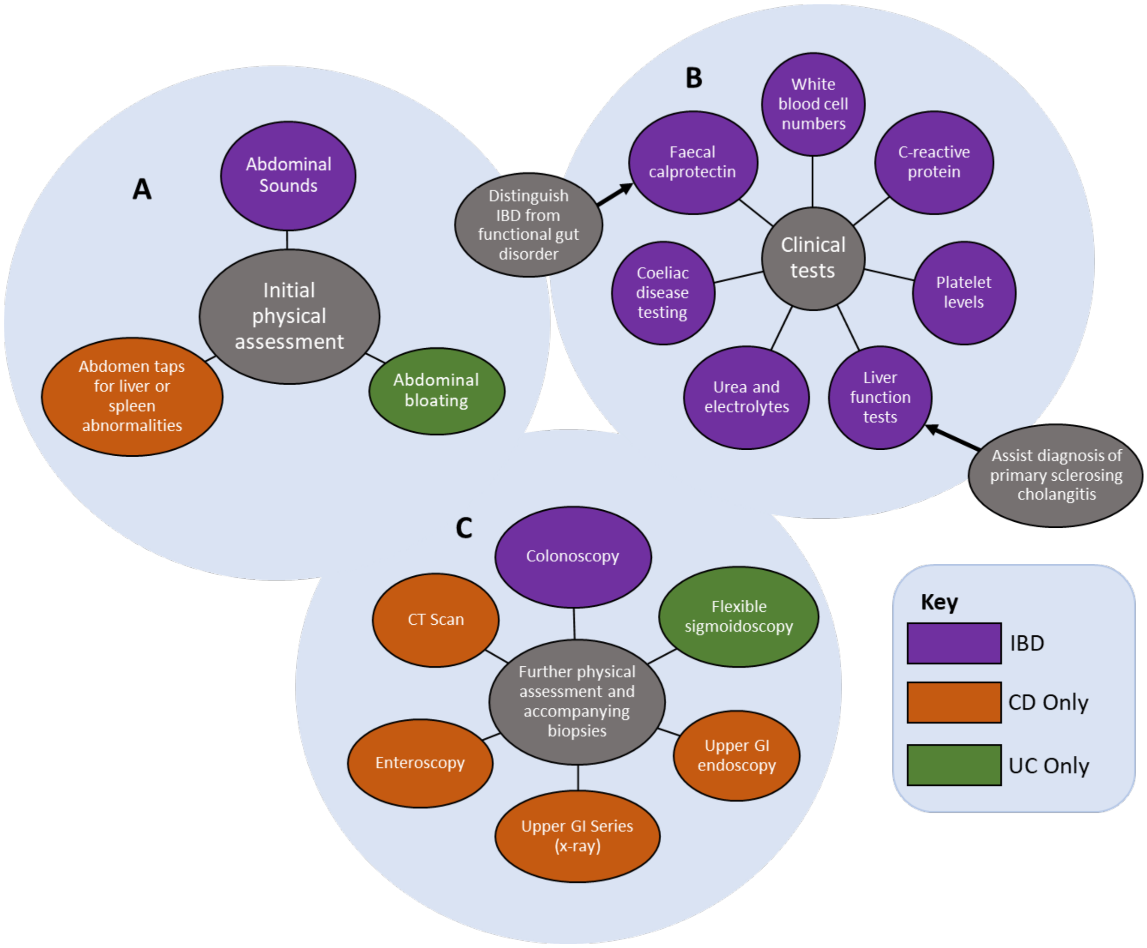


Figure 34 Potential investigations conducted in order to diagnose IBD, and then CD and UC [424-427]. Investigations are colour coded according to whether they are performed for both subtypes, or one subtype. A) Initial physical exams often involve listening to, and feeling the abdomen; B) Clinical tests including blood and stool tests consisting of general inflammatory markers such as platelet count and C-reactive protein, and other tests which can differentiate IBD from other diseases (faecal calprotectin, coeliac disease testing); C) Further physical exams consist of endoscopies where biopsies are taken for histological confirmation of CD or UC. As CD can cause inflammation anywhere in the GI tract, further tests such as enteroscopies are used to investigate disease extent.

Perhaps even more important than an initial diagnosis of CD or UC is an understanding of individual patient’s disease courses. Traditional treatment of IBD involves a step-up approach, reserving the use of more aggressive therapies, for example biologic agents, for severe disease courses, or patients with disease resistant to remission (Section 1.1.6) [428]. An alternative, which is currently being explored, is a top-down approach to treatment. A concise “workflow” for top-down treatment is not currently available, but it broadly suggests that treatments should be given in the inverse order, beginning with biologics, followed by immunomodulators, steroids and lastly 5-ASAs [429]. Few studies have interrogated the alternative top-down approach, although early evidence suggests this is a plausible strategy, and that early administration of a combination of immunosuppressants and biologics were effective at reducing the risks of complications requiring surgery, and increasing time in remission [37]. In a literature review of the efficacy of top-down therapy for Crohn’s disease, the majority of studies were randomised control trials and retrospective cohort studies. The application of the top down therapy was a blanket approach for a subset of a cohort, not a targeted approach [429]. A concern of the top-down approach is the increased risk of adverse reactions in patients when using these more aggressive treatments [429]. The development of methods that can stratify patients based on the severity of disease course could help decide whether a top-down or step-up approach is the most effective form of treatment strategy on a case by case basis.

There are complications specifically associated with CD that can cause irreversible bowel damage, including strictures, fistulas and abscesses [37]. Strictures, or narrowing, can occur in any section of the luminal gastrointestinal tract [430]. They are not uncommon, with around a third of CD patients developing stricture(s) in the first 10 years of their disease course [431]. Studies have demonstrated that patients in early stages of their disease course (less than 2 years) have less bowel damage, and that bowel damage at diagnosis is associated with an increased risk of surgery [37]. This highlights a need to quickly identify cases of CD and treat appropriately to avoid complications and irreversible damage. Therefore, the focus of this chapter is both on IBD subtype diagnosis, and identification of patients who are susceptible to stricturing endotype for early intervention.

Supervised machine learning is an ideal tool for stratification in these cases. There have been several attempts to predict aspects of IBD prognosis, including hospitalisation [253], response to treatment and remission [261, 432], and likelihood of surgical intervention [433]. These types of models almost exclusively use clinical and laboratory data. If these data have to be collected over an extended period of time, it could potentially slow down the rate at which an intervention can be made. Additionally, clinical data such as C-reactive protein and platelet count, which are general measures of inflammation, can be affected by patient co-morbidities, treatments, surgery, and other factors unrelated to a patient’s IBD. This is in contrast to genomic data, which is unaffected by these aforementioned factors. In addition, genomic data remains the same, regardless of the amount of time that has passed since diagnosis, or a patient's current disease status.

Of the recent literature that has sought to combine genetic data and machine learning for IBD, some researchers combined clinical data with specific gene polymorphisms to model early intestinal resection, and extra-intestinal manifestations [434, 435]. Other analysis utilised immunochip genotyping data to classify individuals as CD or controls [436, 437], UC or controls [436], and to assemble a CD risk model that also incorporated clinical information [438]. Earlier work that utilised WES data was published as a response to the Critical Assessment of Genome Interpretation (CAGI) challenge, for classification of CD patients and controls [439]. The datasets associated with the challenge were relatively small, and two of the three datasets had batch effect issues. Many challenge participants chose to select SNVs as features for this challenge. More recently, WES data has been summarised into gene mutational burden scores for classification of CD patients and controls: Wang et al. utilising predicted variant consequence (indel, missense etc.) and zygosity to construct scores [440], and Raimondi et al. used variant consequence, and weighted genes according to their number of appearances in publications where that gene was associated with IBD [441]. A thorough search of the literature reveals no research paper that employs whole exome sequencing in conjunction with machine learning to distinguish between IBD subtypes, or to answer any prognostic questions.

In this chapter, a random forest machine learning algorithm is used for two classification tasks: firstly, to classify patients as the IBD subtype CD, or UC; secondly to determine whether CD patients will develop a stricturing endotype; and thirdly to investigate the impact of age of onset on the genomic basis of IBD. For each classification task three gene panels are used: 1) all genes where GenePy scores are available; 2) an autoimmune gene panel; and 3) an IBD gene panel. Additionally, two different GenePy matrices are used to determine whether an additional filter based on predicted gene pathogenicity is advantageous. Finally, a remapped false positive gene list from [442] is tested as an additional filter for genes included in the random forest model to determine if this additional GenePy matrix pre-processing step improved modelling.

5.2 Methods

5.2.1 Patient phenotype data extraction and characterisation

Patients were recruited according to Section 3.3.1. Diagnoses of IBD subtypes CD, UC and IBDU were made according to British Society of Gastroenterology guidelines for adults [8], and the modified Porto criteria for paediatric patients [369]. Adult IBD subtype diagnosis data was updated as part of preparation of the clinical data. Patients can be diagnosed with IBDU on recruitment, and later their diagnosis is updated to one of the subtypes. Similarly, patients can be mis-diagnosed with UC if their inflammation is only colonic when presenting at clinic, and subsequently inflammation spreads to other areas of the gastrointestinal tract. To confirm each patient's IBD subtype manually, clinical questionnaire information was extracted. If the latest records showed patients had a Harvey Bradshaw Index score on record, the patient was recorded as having CD. If a patient had a recorded Ulcerative Colitis Disease Activity Index, then they were a confirmed UC patient.

Current NHS databases are not automated to provide flags for specific CD endotypes such as stricturing. Additionally, there is no questionnaire or score associated with this endotype. This makes extracting deep phenotyping data more challenging. For paediatric IBD patients, this data had been curated by searching through individual clinic letters. This process was time consuming. Additionally, the increase in sample size comes from patients that are all adults, so on average these patients have more clinical history to search through. In order to extract the adult patient's stricturing endotype status more smoothly to use as an outcome in machine learning, collaborators at the National Institute of Health Research Southampton Biomedical Research Centre assisted in gathering this data. Relevant radiology reports for recruited IBD patients were extracted. These were: endoscopy, small bowel MRI, MR Enterography, abdominal ultrasound, and computerised tomography abdomen scan. To facilitate identification of these endotypes, each report was flagged for presence (1) or absence (0) of keywords related to a stricturing endotype: “strictur”, “fibrosis”, “fibrotic”, “narrowing”, “narrowed”, “dilatation”, “dilati", “stenotic”, and “diameter”. The searches were not case sensitive.

To assess how accurate the flags were, reports for a subset of patients were given to a medical student, with the flags removed. Without knowledge of the flag, the opinion of the medical student and the keywords were assessed for concordance. The flag and the medical student agreed 81% of the time (34/42). In this initial test many false positives were found, as it was common to find endoscopy reports where clinicians had commented “no sign of stricturing”, and these were flagged as positive for the stricturing endotype. In order to assess the instance of false negatives, a second subset of patients for which all reports were flagged 0 in all categories were reviewed blindly by the medical student. In this case, 100% of those reports flagged 0 were confirmed by the medical student as not stricturing. From this analysis, it was concluded that records flagged 0 did not need to be assessed manually and these could be automatically classified as not stricturing. However, due to the number of false positives, all records flagged 1 in any category needed to be manually reviewed. In these initial tests, some reports referred to earlier patient tests which were not present in the dataset. As a result of this, the initial report extraction was widened to ensure reports from all Southampton hospitals were in the dataset. Of 2,398 reports extracted for 506 adult IBD patients, 1,545 were not flagged as 1 for stricturing, leaving 853 reports to review manually. As well as recording the presence of a stricturing endotype, the date of the medical exam when this was first referenced was recorded as the date of the endotype occurrence so that time to stricturing could be calculated for use in further analysis (see Chapter 7).

5.2.2 Additional patient data curation

The patient dataset includes the outcome data for the machine learning tasks (disease subtype, stricturing endotype), but it also contains other important patient information, some of which is used in pre-processing prior to random forest classification. Collating this data involves using the Peddy software [443] and extracting data from BC|INSIGHT. Peddy generates relatedness, IBS0, heterozygosity, sex and ancestry information from a ped file and approximately 25,000 sites of a cohort VCF file. It also compares the content of the ped and VCF files to look for sex mismatches. BC|INSIGHT is the repository for the clinical information collected as part of the Southampton Genetics of IBD study. Deep longitudinal information is collected as part of the study and includes:

i) Demographic information such as sex and date of birth.

ii) Diagnostic information such as date of diagnosis, diagnostic subtype (UC, CD, IBDU), and Paris classification information including age category, extent and severity of each patient’s disease.

iii) Colonoscopy and Gastroscopy: detailed breakdown of the status of areas visualised in colonoscopy and gastroscopy. The reason (for example initial investigation or surveillance) and date of each procedure is recorded.

iv) Details of both autoimmune and non-autoimmune comorbidities.

v) Longitudinal blood test results including reported values for C-reactive protein, platelet count, white blood count and calprotectin.

vi) Surgery information, including surgery type and priority.

vii) Pharmacy data regarding drugs administered, and their dose and frequency.

viii) Anthropometric Data.

5.2.3 Whole exome sequencing data processing

WES data was quality controlled and processed according to Sections 3.3.3 and 3.4, respectively. Two versions of the GenePy matrix were created. The first followed the process according to Section 3.4.2. The second employed an additional filter on the variants included in the GenePy matrix. Only variants that were annotated with a Phred-scaled CADD score ≥ 15 were included in this matrix. Variants with a score of 15 or above would be in the top 3% (approx.) of all possible variation in terms of potential pathogenicity. This threshold has previously been used in the filtering of variants to identify possible disease-causing variation [444, 445]. Therefore, there was a GenePy matrix with all variants included, referred to throughout as **GenePy (all variants)**; and a matrix with variants that had a Phred-scaled CADD score ≥ 15, referred to as **GenePy (CADD cut-off)**.

5.2.4 GenePy score pre-processing

GenePy scores with no variation were excluded using scikit-learn's [446] VarianceThreshold (threshold=0). The remaining scores were scaled by the maximum score of each gene (MaxAbsScaler, disease subtype classifier), or to between 0 and 1 (MinMaxScaler, stricturing endotype classifier) to ensure no bias in downstream machine learning caused by different scoring scales across genes. A further pre-processing step was trialled in order to see if machine learning modelling results could be improved through its implementation. In 2012, Fuentes Fajardo et al. assembled a list of genes which were thought to give a false positive signal in in-silico genomic diagnostics [442]. Reasons for inclusion on this list (hereafter referred to as the **Fuentes false positive gene list**), were highly polymorphic genes, or characteristics that suggested that variants within these genes were miscalls due to technical noise during sequencing. It was discovered that many of the gene symbols listed in the paper were outdated, therefore the gene list was re-mapped. This was achieved by employing the following tools and databases: Multi-

Symbol-Checker from the HUGO Gene Nomenclature Committee (HGNC) [447], g:Profiler [448], Genecards [449, 450], and the NCBI gene database [451, 452]. Once the gene list was remapped, genes in the GenePy matrix that were also present on the Fuentes false positive gene list were filtered out. This additional filter was present for one machine learning modelling pass of the disease subtype and stricturing endotype classifiers, so its effect on the results of machine learning could be assessed.

5.2.5 Patient data pre-processing

Identified by Peddy [443], the most frequent ethnicity was European, so only these cases were included. Only one ethnicity was included to reduce bias in the machine learning modelling. Additionally, there had to be sufficient confidence in the assigned ethnicity, therefore only patients with a probability greater than 90% that the predicted ethnicity was correct were included. Related patients were also removed. For every pair of related individuals, the patient with the younger age of diagnosis was retained for the analysis. The younger patient was included as genetics was more likely to substantially contribute to their IBD aetiology.

5.2.6 Random forest classification

A random forest algorithm was used to perform binary classification tasks for IBD subtypes and the stricturing endotype in Python (v.3.7) using scikit-learn [446]. The model was applied to three different gene panels: 1) all genes with GenePy scores; 2) an autoimmune gene panel curated by HTEdgeSeq; 3) an IBD gene panel that included genes identified in IBD GWAS, and genes associated with monogenic forms of IBD. After genomic, and clinical data pre-processing, the dataset was split into training and testing datasets in an 80:20 ratio, where the split calculation was based on the minority class (UC, patients with stricturing behaviour). Feature selection was performed using a linear support vector classifier (SVC) with L1 penalisation (C=1) using the training data. Cross-validation was used with this feature selection, and the number of folds varied according to the sample size of the training dataset: 10-fold cross-validation for the disease subtype classification, and 5-fold cross-validation was used for the stricturing endotype classifier. Genes not chosen by the classifier in any cross-validation fold of the SVC were excluded. L1, or LASSO, was chosen for feature selection as this method shrinks feature coefficients to zero, essentially removing those features from the dataset. This is in contrast to L2, or ridge penalisation, which retains the features, and makes their associated coefficient very small. LASSO penalisation is ideal for reducing dataset dimensionality, which is necessary for this dataset.

The random forest classifier was trained on the training data using the selected genes. All random forest hyperparameters were set to the default value, aside from the number of estimators (trees), which was set to 10,000. Out-of-bag-error measured the random forest performance on the training data. The random forest ML model was applied to the test set, and its performance evaluated using precision, recall, specificity, F1 score and AUC. Another output was the list of genes chosen in model training ranked by their relative importance to the classifier. The machine learning pipeline can be viewed in Figure 35 (see Supplementary Files for machine learning scripts).

After the most appropriate GenePy matrix and pre-processing steps had been determined, the ML pipeline from Figure 35 was utilised to determine whether differences existed in the genomic basis of IBD depending on the age of onset. To achieve this the IBD dataset was split into each subtype, and the RF classifier attempted to classify CD patients based on whether their disease was paediatric onset (<18 years of age at diagnosis) or adult onset (18 years and over at diagnosis), and then repeated this for the UC patients. Training and testing data for the CD data and UC data were split in an 80:20 ratio, where the split calculation was based on the minority class (adult-onset IBD for both CD and UC data) The same three gene sets were utilised: 1) all genes with GenePy scores; 2) an autoimmune gene panel curated by HTEdgeSeq; 3) an IBD gene panel that included genes identified in IBD GWAS, and genes associated with monogenic forms of IBD.

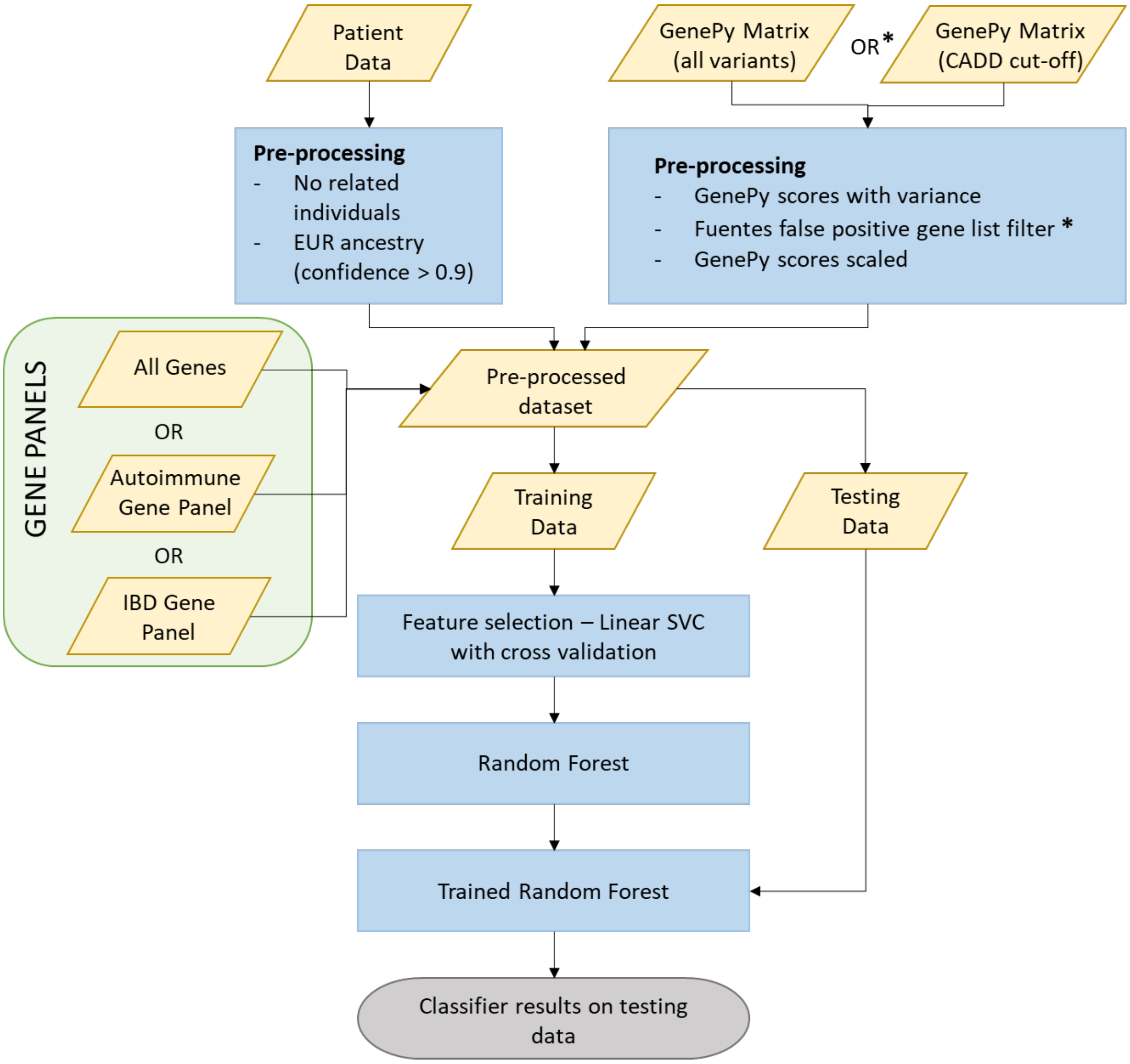

Patient Data
GenePy Matrix (all variants)
OR *
GenePy Matrix (CADD cut-off)
Pre-processing
- No related individuals
- EUR ancestry (confidence > 0.9)
Pre-processing
- GenePy scores with variance
- Fuentes false positive gene list filter *
- GenePy scores scaled
GENE PANELS
All Genes
OR
Autoimmune Gene Panel
OR
IBD Gene Panel
Pre-processed dataset
Training Data
Testing Data
Feature selection – Linear SVC with cross validation
Random Forest
Trained Random Forest
Classifier results on testing data

Figure 35 ML pipeline for classifying IBD subtypes and the CD stricturing endotype. The steps marked with an asterisk (*) indicate the two places where different data or processing was implemented. ML results with and without these changes were analysed to see how this affected the ML model performance.

5.2.7 Analysis of selected features

After obtaining model results from best performing subtype and stricturing endotype classification ML models, SHAP values were used to gain further insights into how genes contributed to these classifications. SHapley Additive exPLanations, or SHAP values, are an explainable AI tool based on the mathematical concept of the Shapley value, which measures the average marginal contribution of each variable [453]. This analysis was performed in Python using the SHAP package, specifically the SHAP tree explainer tool.

Pathway enrichment analysis was generated for the model with the highest AUC for both the disease subtype and stricturing endotype classifiers. Genes selected by the Linear SVC feature selection during operation of the machine learning pipeline were input into the Enrichr [454]. The maximum number of genes was included in all instances of analysis (there is a maximum threshold of top relevant genes to include in analysis of 500). Significantly enriched pathways were determined using the KEGG [455] 2021 Human database. Pathways were determined to be enriched according to the Fisher exact test p-value < 0.05, adjusted using the Benjamini-Hochberg multiple hypotheses testing correction (as is standard for the Enrichr software).

5.3 Results

The IBD cohort includes 1,087 recruited individuals that have been whole exome sequenced. Of these, 502 patients were recruited in the paediatric clinic, and 506 were recruited in the adult clinic. The remaining 79 were parents or relatives of the proband that was initially recruited, some of whom also have an IBD diagnosis. Table 15 characterises the cohort further. It is important to note that some patients recruited through the adult IBD clinic may have been diagnosed as children.

Table 15 Clinical characteristics of the IBD cohort, split by paediatric and adult IBD diagnosis. Some categories do not sum to 1,087 because of incomplete data.

		Paediatric IBD (<18 years)	**Adult IBD (≥18 years)**	**Total**
Age at diagnosis	Median years (N, range)	13 (577, 1-17)	32 (496, 18-84)	1,073
Subtype diagnosis	CD	379	297	676
	UC	176	190	366
	IBDU	20	6	26
Stricturing Endotype (CD Only)	Yes	77	103	180
	No	480	324	804
Sex	Male	328	226	554
	Female	249	270	519
Ancestry (Peddy)	African (AFR)	3	4	7
	American (AMR)	8	1	9
	East Asian (EAS)	0	1	1
	European (EUR)	545	471	1,016
	South Asian (SAS)	13	13	26
	Unknown	8	5	13

As detailed in Section 3.3.3, quality control was performed on the WES data of this IBD cohort. One sample, which had already been sent for sequencing previously, was found to mismatch with the SNP fingerprinting performed. That sample was excluded, but it did not impact the overall number of individuals in the cohort as it was a duplicate. Aside from this, the checks performed showed no other mismatches or contaminations, and as such all other samples passed quality control checks (see Supplementary Files). The characteristics of the cohorts used for 1) the IBD subtype classifier, and 2) the stricturing endotype classifier, after patient data pre-processing steps regarding ancestry prediction and relatedness were completed, are found in Table 16 and Table 17, respectively.

Table 16 Clinical characteristics of individuals included in the IBD subtype classifier models, after patient data pre-processing. Age at diagnosis information was unavailable for three individuals.

		Paediatric IBD (< 18 yrs)	**Adult IBD (≥18 yrs)**
N		491	412
Median age at diagnosis (range)		13 (1-17)	32 (18-82)
IBD Subtype	CD	334	263
	UC	157	149
Sex	Male	286	191
	Female	205	221

Table 17 Clinical characteristics of individuals included in the CD stricturing endotype models, after patient data pre-processing. Age at diagnosis information was unavailable for two individuals.

		Paediatric IBD (< 18 yrs)	**Adult IBD (≥18 yrs)**
N		332	255
Median age at diagnosis (range)		13 (1-17)	31 (18-82)
Stricturing Endotype	Yes	71	98
	No	261	157
Sex	Male	206	113
	Female	126	142

5.3.1 Impact of different GenePy matrix formulations on random forest modelling

In total, 335,978 exonic variants were input into the GenePy (all variants) matrix. For the GenePy (CADD cut-off) matrix 135,867 exonic variants had a Phred-scaled CADD score ≥ 15. An increased number of variants does give more power to detect causal variants in each patient, but raises a potential concern when integrating this information into a GenePy score. The GenePy score for some genes may become artificially inflated because of the summation of many variants with minimal effect. It is for this reason that the GenePy (CADD cut-off) matrix was generated, where fewer variants with a larger potential effect size would be included in each score.

The machine learning pipeline process was repeated three times for the two GenePy score datasets. Each pipeline run uses a different gene panel: 1) all available genes; 2) an autoimmune gene panel curated by HTGEdgeSeq; 3) an IBD gene panel curated in house, including genes implicated in GWAS and genes reported as causing monogenic forms of IBD (see Supplementary files for gene panels). Whether a GenePy score is available for a gene is dependent on two factors: if that gene can be annotated by Ensembl-VEP [456], and if there are variants in the cohort in that gene that met the rigorous quality filters. A breakdown of the number of genes included in downstream machine learning after each pre-processing stage is detailed in Table 18. The comparison of the two GenePy score matrices was completed with two classification tasks 1) the disease subtype classifier discriminating CD and UC, and 2) the stricturing endotype classifier on CD patients only. A breakdown of the training and testing datasets for each classification task is detailed in Table 19.

Table 18 Number of genes with GenePy scores for the GenePy (all variants) matrix, and GenePy (CADD-cut-off) matrix before and after selecting genes with GenePy score variance. This breakdown is shown for every gene panel, and for both the disease subtype classifier, and the stricturing endotype classifier.

			Total genes with GenePy scores	**Genes with GenePy score variance (% of total)**
Subtype Classifier (n=906)	GenePy (all variants)	All genes	16,794	16,657 (99.2%)
		Autoimmune gene panel	1,721	1,706 (99.1%)
		IBD gene panel	526	523 (99.4%)
	GenePy (CADD cut-off)	All genes	15,669	15,341 (97.9%)
		Autoimmune gene panel	1,598	1,552 (97.1%)
		IBD gene panel	499	494 (99.0%)
Stricturing Classifier (n=589)	GenePy (all variants)	All genes	16,794	16,465 (98.0%)
		Autoimmune gene panel	1,721	1,692 (98.3%)
		IBD gene panel	526	518 (98.5%)
	GenePy (CADD cut-off)	All genes	15,669	14,742 (94.1%)
		Autoimmune gene panel	1,598	1,493 (93.4%)
		IBD gene panel	499	472 (94.6%)

Table 19 Training and testing dataset sizes for the disease subtype classifier and the stricturing subtype classifier

	Training Dataset		**Testing Dataset**		**Total**
	CD	*UC*	*CD*	*UC*	
Disease Subtype Classifier	244	244	356	62	906
	Stricturing	*Not Stricturing*	*Stricturing*	*Not Stricturing*	
Stricturing Subtype Classifier	136	136	34	283	589

The results of the disease subtype classifier on the test dataset for both GenePy datasets are contained in Table 20. Regardless of the gene panel used as input for machine learning, using the GenePy matrix with the CADD cut-off produces higher AUC and F1 scores for classifying CD and UC on the test data. This gives some evidence that the CADD cut-off is beneficial, as it may reduce the noise in patients GenePy scores caused by many low-effect size variants. Instead the GenePy scores for each gene sum together fewer variants with a larger predicted pathogenic effect. When analysing the results from models using the GenePy matrix with the CADD cut-off, the best model uses the autoimmune gene panel, achieving an AUC of 0.67. In comparison to the other two models, an uplift in sensitivity to UC cases is observed (sensitivity 0.58 versus 0.53 and 0.52). For every gene panel, and the different GenePy matrices, *NOD2* is present as the strongest genetic signal for all random forest models. This is not surprising, given the known potential impact of *NOD2* variants on the development of CD. It is reassuring to see the machine learning model identify this gene as the strongest discriminant. Aside from *NOD2*, two genes remain constant to the top 10 for the all genes classifiers, and three genes for the autoimmune and IBD panels. These are: *ASPM* and *EPB41L4A* for the all genes classifier; *DNAH12*, *TNS1* and *HTT* for the autoimmune gene panel; and *NFATC1*, *ERAP1* and *DOCK8* for the IBD gene panel.

A comparison of the distribution of *NOD2* GenePy scores can be viewed in Figure 36, where the results from the best model using the GenePy with all variants, and GenePy matrix using the CADD cut-off, are compared. There is a strong difference in *NOD2* distribution depending on which matrix is used, as *NOD2* is more important in the classifier where the CADD cut-off is used than where all variants were used. Many of the other gene distributions were similar when comparing the CD and UC classes, even when utilising the GenePy matrix with CADD cut-off. Often one of the disease subtype classes will have a longer tail to the distribution, i.e. a few individuals are present in the subtype class with a high GenePy score.

Table 20 Random forest classifier of UC and CD. Machine learning metrics on the test set for both versions of GenePy scores, and three different feature sets

GenePy (all variants) – all genes					**GenePy (all variants) – autoimmune gene panel**					**GenePy (all variants) – IBD gene panel**				
No. Features	1,240				No. Features	826				No. Features	459			
	Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1
CD	0.87	0.53	0.53	0.66	CD	0.88	0.54	0.58	0.67	CD	0.87	0.57	0.52	0.69
UC	0.17	0.53	0.53	0.25	UC	0.18	0.58	0.54	0.28	UC	0.17	0.52	0.57	0.26
Average	0.76	0.53	0.53	0.60	Average	0.78	0.55	0.58	0.61	Average	0.77	0.56	0.52	0.63
AUC	0.53				AUC	0.64				AUC	0.57			
Top 10 Genes	*NOD2, ASPM, TAPBPL, BOD1L1, SPTBN5, USP40, EPB41L4A, GRIN2B, POMT2, SYNE1*				Top 10 Genes	*NOD2, MEFV, CX3CR1, DNAH12, TNS1, NCOR2, P2RX7, HTT, TSHR, ERAP1*				Top 10 Genes	*NOD2, MEFV, NFATC1, PER3, ERAP1, TNFRSF6B, DOCK8, ADA2, BANK1, TAF8*			
GenePy (CADD cut-off) – all genes					**GenePy (CADD cut-off) – autoimmune gene panel**					**GenePy (CADD cut-off) – IBD gene panel**				
No. Features	1,213				No. Features	733				No. Features	403			
	Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1
CD	0.88	0.63	0.50	0.73	CD	0.91	0.62	0.66	0.74	CD	0.87	0.58	0.52	0.70
UC	0.19	0.50	0.63	0.27	UC	0.23	0.66	0.62	0.34	UC	0.18	0.52	0.58	0.26
Average	0.78	0.61	0.52	0.66	Average	0.81	0.62	0.65	0.68	Average	0.77	0.57	0.53	0.63
AUC	0.59				AUC	0.67				AUC	0.59			
Top 10 Genes	*NOD2, GC, EPB41L4A, ASPM, LAMA1, VWDE, COL4A3, TUBB3, DNAH17, SVEP1*				Top 10 Genes	*NOD2, TTN, TG, DNAH12, TNS1, P2RX7, WDFY4, TNC, SPATS2L, HTT*				Top 10 Genes	*NOD2, GC, DOCK8, NPC1, GALC, ERAP1, NFATC1, CELSR3, TEP1, CD6*			

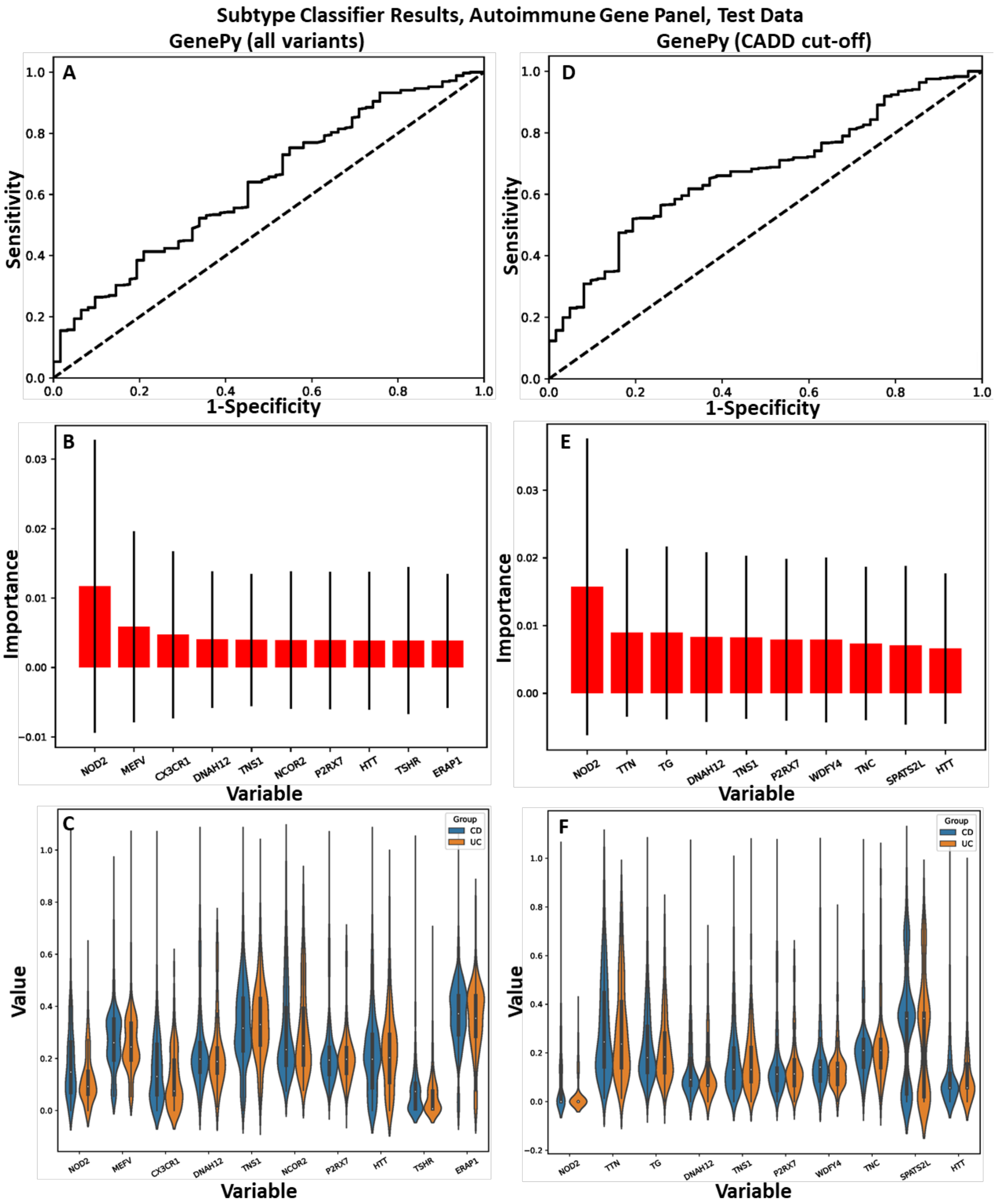


Figure 36 Comparison of the best disease subtype model produced using a GenePy matrix generated with all variants (A-C) and GenePy with variants that meet the CADD cut-off (D-F). In both cases the random forest model performed best with the autoimmune gene panel. A) AUC on the test set for GenePy (all variants); B) Top 10 most discriminate genes and their relative importance in the GenePy (all variants) random forest; C) Violin plots of the top 10 most discriminant genes (CD=blue,

UC=orange) for GenePy (all variants); D) AUC on the test set for GenePy (CADD cut-off); E) Top 10 most discriminate genes and their relative importance in the GenePy (CADD cut-off) random forest; C) Violin plots of the top 10 most discriminant genes (CD=blue, UC=orange) for GenePy (CADD cut-off).

The results for the stricturing endotype classifier on the test data can be viewed in Table 21. In this case it is also seen that the classification results are the same or better across the gene panels when utilising the GenePy matrix with the CADD cut-off. In particular, the random forest classifiers that used all genes, and the IBD panel, saw an AUC increase of 0.1 after implementation of the CADD cut-off. This is reflected in a complete change in the 10 most important genes to classification for the ML model that used all genes, and only the genes *CNTRL* and *GC* are present in both classifiers that use the IBD panel. For the two classifiers that use the autoimmune gene panel, the top 10 genes in both are completely different, but this is not surprising for a ML model where the performance is no better than random (AUC 0.5 for both versions of the classifier). A comparison of the results of the best ML model (all genes, AUC 0.59) for the GenePy matrix with all variants, and the GenePy matrix with the CADD cut-off is presented in Figure 37. There is very little difference between the GenePy score distributions of the top 10 genes for the stricturing and not-stricturing classes, in both ML models shown in Figure 37. Of surprise here is an absence of *NOD2* in all top 10 gene lists apart from the classifier that used the IBD gene panel. *NOD2* is also implicated in the formation of strictures, and variation in this gene has been shown to be a risk factor in the development of this endotype [5]. Overall, these results provide evidence that utilising the CADD cut-off when generating the GenePy matrix is beneficial for downstream machine learning modelling.

Table 21 Random forest classifier of stricturing (S) vs not-stricturing (NS) in CD patients. Machine learning metrics on the test set for both versions of GenePy scores, and three different feature sets

GenePy (all variants) – all genes					**GenePy (all variants) – autoimmune gene panel**					**GenePy (all variants) – IBD gene panel**				
No. Features	570				No. Features	467				No. Features	317			
	Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1
NS	0.90	0.46	0.59	0.61	NS	0.9	0.50	0.56	0.65	NS	0.86	0.44	0.41	0.58
S	0.11	0.59	0.46	0.19	S	0.12	0.56	0.50	0.20	S	0.08	0.41	0.44	0.14
Average	0.82	0.47	0.57	0.56	Average	0.82	0.51	0.55	0.60	Average	0.78	0.44	0.41	0.54
AUC	0.488				AUC	0.50				AUC	0.44			
Top 10 Genes	*FABP2, EPN3, SEC16A, HECW1, TOMM34, PTPRQ, ABCC6, C4orf50, TNN, THSD7B*				Top 10 Genes	*TPO, SEC16A, FABP2, PADI4, CARD14, GPR35, NOTCH1, NCOR2, KSR1, ANK3*				Top 10 Genes	*SEC16A, CNTRL, GPR35, NOTCH1, KSR1, BANK1, FAM171B, GC, IRF2BP2, RPS6KA2*			
GenePy (CADD cut-off) – all genes					**GenePy (CADD cut-off) – autoimmune gene panel**					**GenePy (CADD cut-off) – IBD gene panel**				
No. Features	520				No. Features	418				No. Features	292			
	Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1
NS	0.92	0.57	0.59	0.70	NS	0.88	0.53	0.41	0.66	NS	0.89	0.47	0.53	0.61
S	0.14	0.59	0.57	0.23	S	0.09	0.41	0.53	0.15	S	0.11	0.53	0.47	0.18
Average	0.84	0.57	0.59	0.65	Average	0.80	0.51	0.42	0.61	Average	0.81	0.47	0.52	0.57
AUC	0.59				AUC	0.50				AUC	0.54			
Top 10 Genes	*PREX1, CNTRL, MAPT, SVEP1, TTN, FAT4, OR5M1, PKD1L3, PLCE1, PTPRQ*				Top 10 Genes	*TG, TTN, TNS1, P2RX7, TNC, LOXL2, SPATS2L, BAZ2B, DNAH12, FLT4*				Top 10 Genes	*GC, CNTRL, DOCK8, UTP20, TEP1, NPC1, F5, GALC, GSDMA, NOD2*			

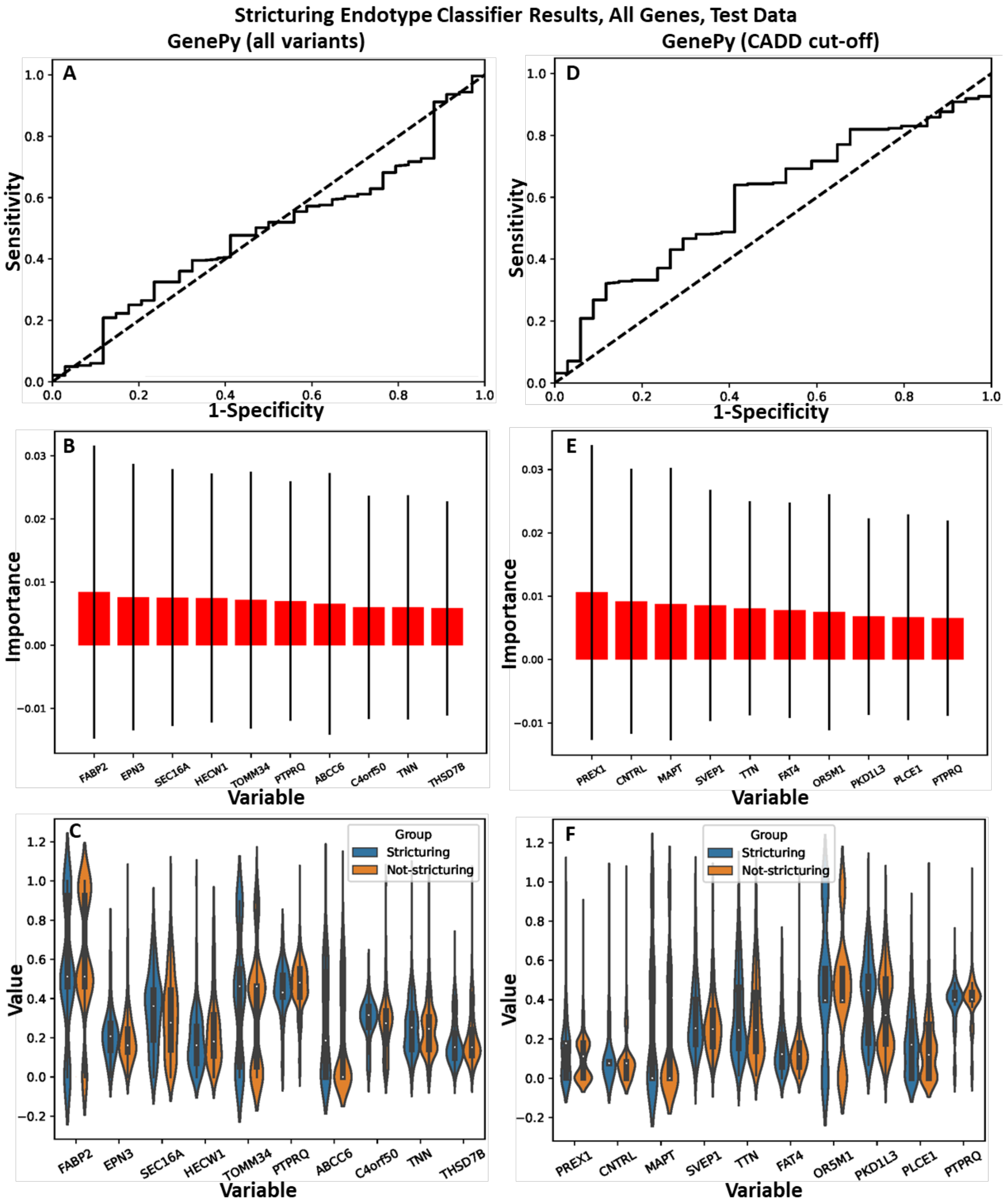


Figure 37 Comparison of the stricturing endotype random forest model produced using a GenePy matrix with all variants (A-C) and a GenePy matrix utilising the CADD cut-off (D-F), using the all genes where GenePy scores were available. A) AUC on the test set for GenePy (all variants); B) Top 10 most discriminate genes and their relative importance in the GenePy (all variants) random forest; C) Violin plots of the top 10 most discriminant genes (stricturing=blue, not-stricturing=orange) for GenePy (all

variants); D) AUC on the test set for GenePy (CADD cut-off); E) Top 10 most discriminate genes and their relative importance in the GenePy (CADD cut-off) random forest; F) Violin plots of the top 10 most discriminant genes (stricturing=blue, not-stricturing=orange) for GenePy (CADD cut-off).

5.3.2 Impact of Fuentes false positive list on machine learning classifier results

The comparison of the random forest classifiers concluded that for these data, using the GenePy matrices constructed with variants with a CADD score ≥ 15 is the best strategy. However, one gene observed in the top 10 genes for the stricturing versus not stricturing classifier raised concerns. In the all genes and autoimmune gene panel stricturing endotype classifier, and the autoimmune gene panel disease subtype panel that used the GenePy (CADD cut-off) matrix, *TTN* is in the top 5 most discriminant genes. This is the longest human gene and as such can accrue many mutations (and therefore a high GenePy score) without this necessarily contributing to disease.

In order to potentially exclude genes such as *TTN*, which are highly mutable, but also highly unlikely to cause disease, the Fuentes false positive gene list was implemented as an additional GenePy score pre-processing step. However, during initial investigations of the Fuentes false positive gene list, of 2,213 genes in the list, 1,644 were found to not be present in GenePy. It was subsequently determined that many of the gene symbols on the list needed to remapped. Using Multi-Symbol-Checker from HGNC [447], or g:Profiler [448] – where an Ensembl gene ID was identified and then converted to a gene symbol – remapping, or confirmation that the original gene symbol was correct, was performed for 1,564 genes. Two genes were identified as being withdrawn from databases.

For the remaining 649 genes, the Genecards database [449, 450] was searched to identify alternative aliases. Additionally, any gene symbols starting with “LOC” were searched without the prefix to find if these genes had been identified. It was established through the NCBI gene database [451, 452] that three groups of unidentified genes (prefixes FLJ, DKFZp, and MGC) were clones of another gene symbol. An additional three withdrawn genes were identified through these database checks. After these searches, another 578 gene symbols had been remapped. In total, 2,141 genes were remapped, or their gene symbol was confirmed. However, there were many duplicates, and after these were removed 1,298 genes remained on the Fuentes false positive gene list. The remapped gene list can be found in the supplementary files. In Table 22, the number of genes present after each GenePy score pre-processing stage can be viewed.

Table 22 Number of genes with GenePy scores in the GenePy (CADD cut-off) matrix at each stage of pre-processing the data prior to machine learning. Also includes the percentage change between the genes with GenePy scores, and the number of genes included after implementation of both pre-processing stages.

		Total genes with GenePy scores	**Genes after exclusion using false positive list**	**Genes with GenePy score variance**	**% Change**
Subtype Classifier (n=906)	All genes	15,669	15,242	14,922	4.8
	Autoimmune gene panel	1,598	1,586	1,540	3.6
	IBD gene panel	499	494	489	2.0
Stricturing Classifier (n=589)	All genes	15,669	15,242	14,342	8.5
	Autoimmune gene panel	1,598	1,586	1,484	7.1
	IBD gene panel	499	494	467	6.4

The results for disease subtype, and stricturing endotype machine learning classifiers that exclude genes on the Fuentes false positive list, are collated in Table 23. For the disease subtype classifier there were very minor changes in the AUC in comparison to the classifier which did not use the Fuentes false positive gene list. After employing this additional filter, the AUC for the classifier that uses all genes reduced by 0.02, with one change in the top 10 important genes (*SVEP1* in previous ML results is replaced by *MYO18B*). The number of genes chosen by feature selection increased by 3, to 1,216. The *TTN* and *TG* genes were replaced with *E2F4* and *NFATC1* for the classifier that began with the autoimmune gene panel, with no change in the AUC, and 6 fewer genes selected during feature selection. The performance of the model that uses the IBD gene panel improves slightly with this additional gene filter, achieving an AUC of 0.6. As with the autoimmune panel classifier, 6 fewer genes are chosen during feature selection. *TEP1* is replaced by *GSDMA* in the top 10 most important genes. Overall, random forest performance is not better or worse for the disease subtype classifier with the use of the Fuentes false positive gene list as a filter. However, it does remove a gene signal from ML modelling that is known to be erroneous. The best ML model results, using the autoimmune gene panel, with the Fuentes filter, are shown in Figure 38.

Table 23 Fuentes results on the test set for both the disease subtype classifier and the stricturing classifier

CD vs UC– ALL GENES					**CD vs UC – AUTOIMMUNE PANEL GENES**					**CD vs UC – IBD PANEL GENES**				
No. Features	1,216				No. Features	739				No. Features	397			
	Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1
CD	0.88	0.62	0.53	0.73	CD	0.90	0.61	0.63	0.73	CD	0.87	0.56	0.52	0.68
UC	0.20	0.53	0.62	0.59	UC	0.22	0.63	0.61	0.33	UC	0.17	0.52	0.56	0.25
Average	0.78	0.61	0.54	0.66	Average	0.80	0.61	0.63	0.67	Average	0.77	0.55	0.56	0.62
AUC	0.57				AUC	0.67				AUC	0.60			
Top 10 Genes	*NOD2, GC, EPB41L4A, ASPM, LAMA1, VWDE, COL4A3, TUBB3, MYO18B, DNAH17*				Top 10 Genes	*NOD2, DNAH12, TNS1, WDFY4, P2RX7, SPATS2L, TNC, HTT, E2F4, NFATC1*				Top 10 Genes	*NOD2, GC, DOCK8, NPC1, GALC, ERAP1, NFATC1, CELSR3, GSDMA, CD6*			
STRICTURING VS NOT-STRICTURING– ALL GENES					**STRICTURING VS NOT-STRICTURING – AUTOIMMUNE PANEL GENES**					**STRICTURING VS NOT-STRICTURING – IBD PANEL GENES**				
No. Features	534				No. Features	411				No. Features	284			
	Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1
NS	0.92	0.54	0.59	0.68	NS	0.90	0.51	0.53	0.65	NS	0.89	0.45	0.56	0.60
S	0.13	0.59	0.54	0.22	S	0.12	0.53	0.51	0.19	S	0.11	0.56	0.45	0.18
Average	0.83	0.54	0.58	0.63	Average	0.82	0.51	0.53	0.60	Average	0.81	0.46	0.55	0.55
AUC	0.63				AUC	0.52				AUC	0.55			
Top 10 Genes	*PREX1, CNTRL, MAPT, FAT4, GC, AKR7L, PLCE1, PKD1L3, ACACB, PTPRQ*				Top 10 Genes	*TNS1, P2RX7, SPATS2L, LOXL2, BAZ2B, DNAH12, ANK3, FLT4, SORBS1, WDFY4*				Top 10 Genes	*GC, CNTRL, DOCK8, UTP20, NPC1, GALC, GSDMA, NOD2, ERAP1, CD6*			

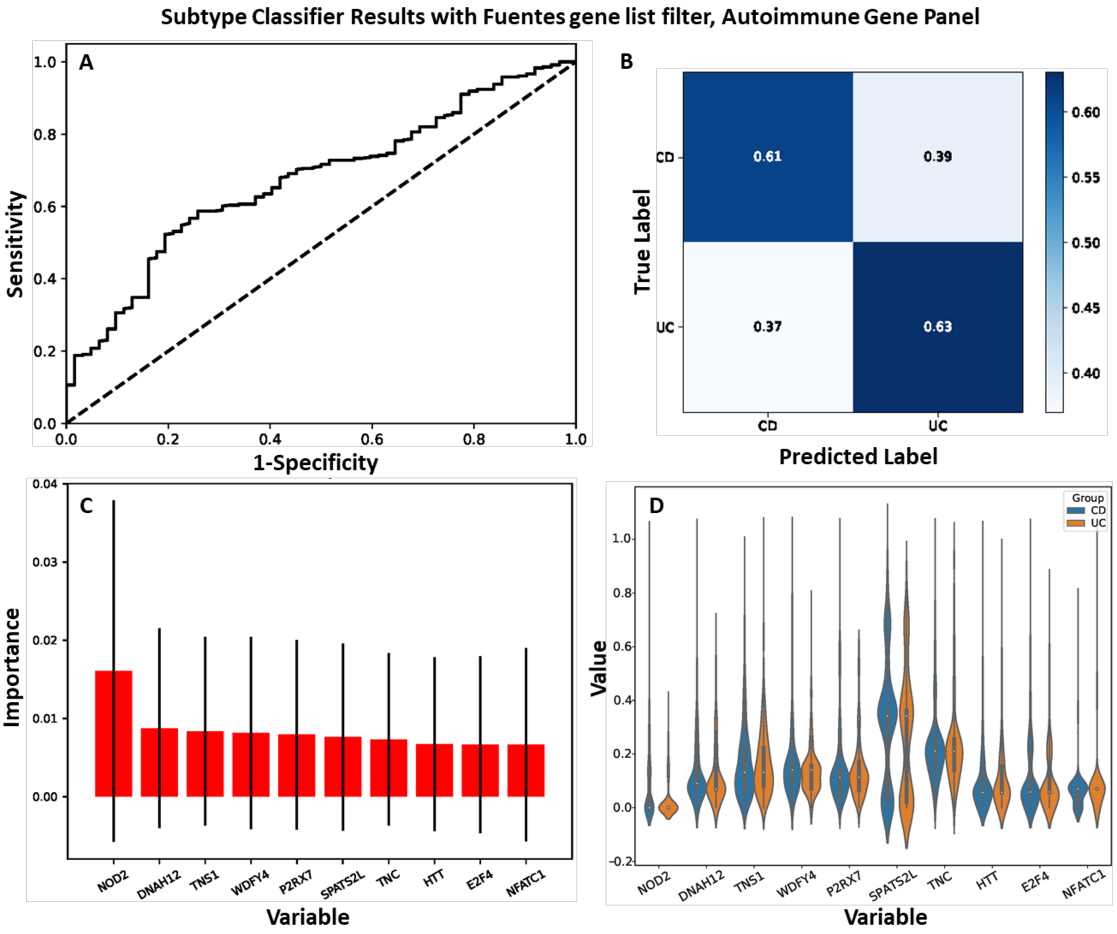


Figure 38 IBD subtype random forest model using the Fuentes false positive gene list as an additional filter, and the GenePy matrix with the CADD cut-off. A) Test dataset AUC; B) Normalised confusion matrix on test dataset. C) Top 10 most discriminate genes and their relative importance for the random forest; D) Violin plots of the top 10 most discriminant genes (CD=blue, UC=orange).

A further examination of the contributions that different genes made to classification of IBD subtypes by the random forest model was conducted by producing SHAP values (Figure 39A). SHAP values revealed that, in general, a low GenePy score contributed to UC classification (a negative SHAP value), and a high GenePy score contributed to CD classification (a positive SHAP value). The genes for which this trend did not apply were *NFATC1*, LRR1, IL31RA, NRP1, PYGL and LRP1. There was also an extended look at the feature importances, as shown in previous figures such as Figure 38C. In Figure 39B, the feature importance value of the top 50 genes (739 genes were selected in total by feature selection, as documented in Table 23) are shown. As the feature importances of all 739 genes sum to 1, and this visualises the ever decreasing contribution that each feature makes to classification.

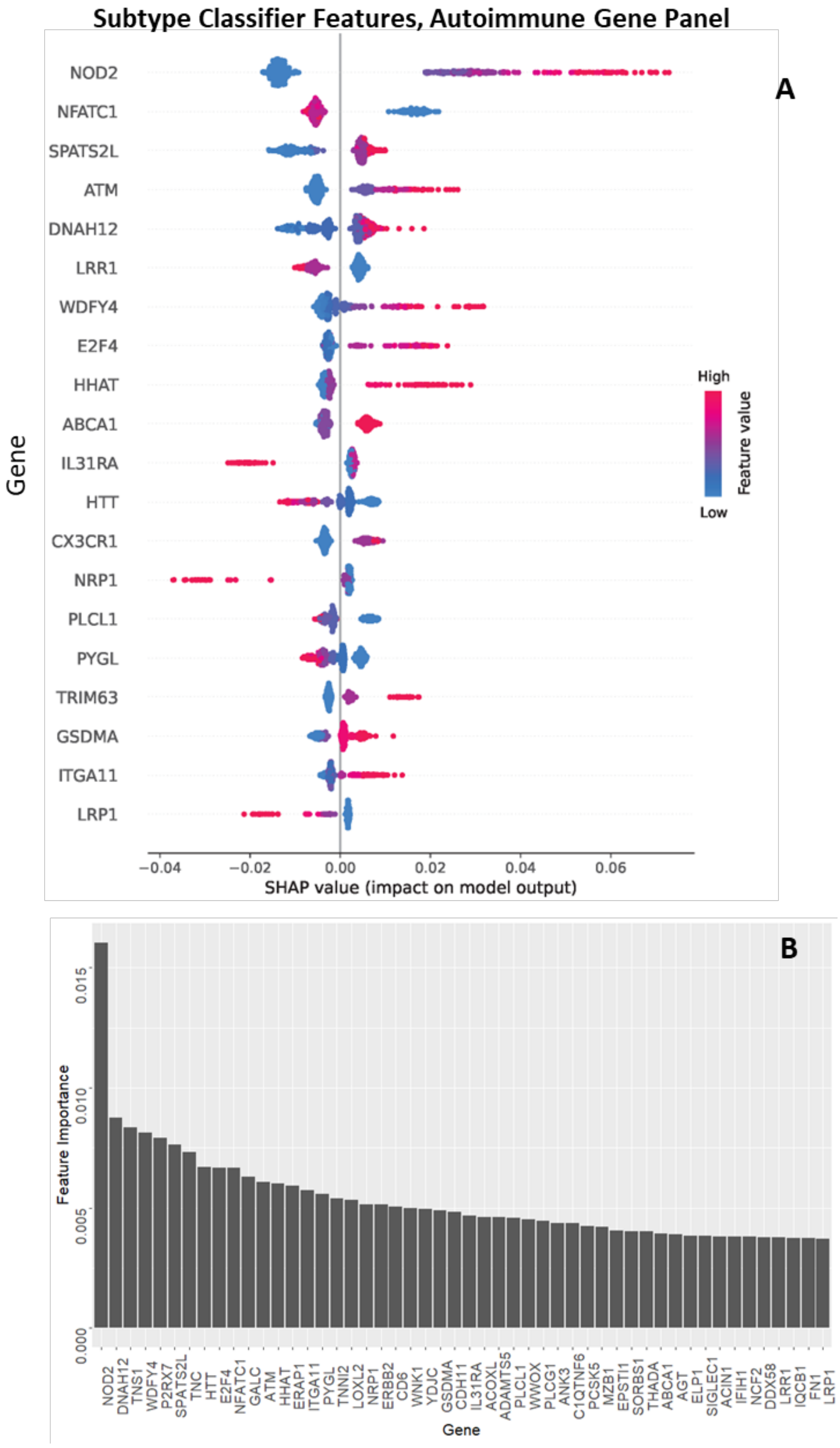


Figure 39 Further analysis of gene contributions to the best subtype classifier, using the GenePy (CADD cut-off) matrix, and the Fuentes filter. A) SHAP values for top discriminatory genes, where a high feature value is equivalent to a high GenePy score and vice versa. A positive SHAP value indicates the feature makes a contribution to the positive class, which was coding as CD. B) Feature importance as in Figure 38C, but extended to the top 50 genes.

In addition, pathway enrichment analysis was performed using Enrichr [454] and the KEGG [455] 2021 Human database, with the genes selected during feature selection for the best performing IBD subtype model, which used the autoimmune gene panel. This produced 289 pathways, of which 162 were significant after adjusting the p-value for multiple hypotheses testing (adjusted p-value < 0.05). Table 24 lists the top 20 most significant pathways according to the combined score produced by Enrichr, which takes into account p-value and the z-score for the deviation from the expected rank. This revealed several immune pathways that were enriched in this gene set, some of which are already know to contribute to IBD aetiology such as the JAK-STAT signalling pathway [457] and the NF-κB signalling pathway [98, 387]. However, there was concern with this approach that the reason these pathways were enriched was due to how the autoimmune gene panel was constructed, as it is naturally enriched for immune pathways. This issue was exacerbated by the number of genes chosen during feature selection, as this meant 48% of the genes present in the autoimmune gene panel (post gene-filtering steps) were input into random forest modelling. Pathway enrichment analysis was performed using the 1,540 autoimmune genes in the panel prior to any feature selection. This produced 303 pathways, of which 204 had a significant adjusted p-value. Aside from two pathways (other glycan degradation, ABC transporters), all significant pathways from the enrichment analysis of genes selected for subtype classification overlapped with significant pathways contained within the whole autoimmune gene panel. This led to the thought that observing which pathways had been removed during feature selection would be more appropriate. Here, it was found that 44 pathways were no longer significant. Of particular interest was the exclusion of the terms type 1 diabetes mellitus, and autoimmune thyroid disease. The full list of pathways omitted by feature selection can be viewed in Supplementary Table 3.

Table 24 Pathways identified by Enrichr as significant (according to an adjusted p-value < 0.05) from the features selected during subtype classifier modelling. Top 20 of 162, ordered by combined score.

Term	**Overlap**	**P-value**	**Adjusted P-value**	**Odds Ratio**	**Combined Score**
PPAR signalling pathway	28/74	1.87E-21	1.35E-19	16.45	785.13
JAK-STAT signalling pathway	46/162	3.84E-28	3.70E-26	10.96	691.58
Th1 and Th2 cell differentiation	29/92	1.32E-19	5.46E-18	12.45	541.05

Term	**Overlap**	**P-value**	**Adjusted P-value**	**Odds Ratio**	**Combined Score**
Cytokine-cytokine receptor interaction	61/295	1.37E-28	1.98E-26	7.32	469.35
Adipocytokine signalling pathway	22/69	2.58E-15	3.74E-14	12.54	421.33
Th17 cell differentiation	29/107	1.35E-17	3.25E-16	10.04	390.19
Pathways in cancer	85/531	6.02E-31	1.74E-28	5.48	381.53
PD-L1 expression and PD-1 checkpoint pathway in cancer	25/89	9.33E-16	1.50E-14	10.50	363.47
AGE-RAGE signalling pathway in diabetic complications	27/100	1.99E-16	4.41E-15	9.97	360.38
C-type lectin receptor signalling pathway	27/104	5.87E-16	1.07E-14	9.45	331.36
NF-κB signalling pathway	27/104	5.87E-16	1.07E-14	9.45	331.36
Tuberculosis	39/180	1.95E-19	7.05E-18	7.56	325.47
Coronavirus disease	46/232	5.05E-21	2.92E-19	6.81	318.14
Hepatitis B	36/162	2.02E-18	5.30E-17	7.78	316.86
Type II diabetes mellitus	15/46	5.00E-11	3.21E-10	12.85	304.84
Hematopoietic cell lineage	25/99	1.43E-14	1.76E-13	9.08	289.43
FoxO signalling pathway	30/131	5.95E-16	1.07E-14	8.03	281.41
Lipid and atherosclerosis	42/215	4.78E-19	1.54E-17	6.65	280.46
Insulin resistance	26/108	1.46E-14	1.76E-13	8.53	271.69
Osteoclast differentiation	29/127	1.98E-15	3.01E-14	7.99	270.40

For the stricturing endotype ML models that employed the Fuentes false positive gene list filter (results Table 23), the AUC improved regardless of the gene panel utilised. For the classifier using all genes the AUC improved by 0.04, and for the autoimmune panel and IBD panel AUC increased by 0.02 and 0.01, respectively (Figure 40). There were small changes to the number of genes chosen in feature selection: 14 fewer for the all genes classifier, 7 fewer when using the autoimmune panel, and 8 fewer genes for the IBD panel. For the top 10 genes in the all genes ML model, *SVEP1*, *TTN* and *OR5M1* are replaced by *GC*, *AKR7L* and *ACACB* after utilising the false positive gene list. Three genes are also replaced for the autoimmune gene panel random forest: *TG*, *TTN* and *TNC* are changed to *ANK3*, *SORBS1* and *WDFY4*. Only two genes change between classifiers for the IBD panel: *TEP1* and *F5* are replaced by *ERAP1* and *CD6*. When comparing the most discriminant genes for the disease subtype classifier and the stricturing endotype classifier, there are some commonalities. For the all genes classifiers, only *GC* is common to both top 10. However, for the autoimmune gene panel there were five shared genes between the two classification tasks (*DNAH12*, *TNS1*, *WDFY4*, *P2RX7*, and *SPATS2L*), and for the IBD gene panel there were seven (*NOD2*, *GC*, *DOCK8*, *NPC1*, *GALC*, *GSDMA*, and *CD6*). This suggests that CD-associated genes are driving the disease subtype classifier.

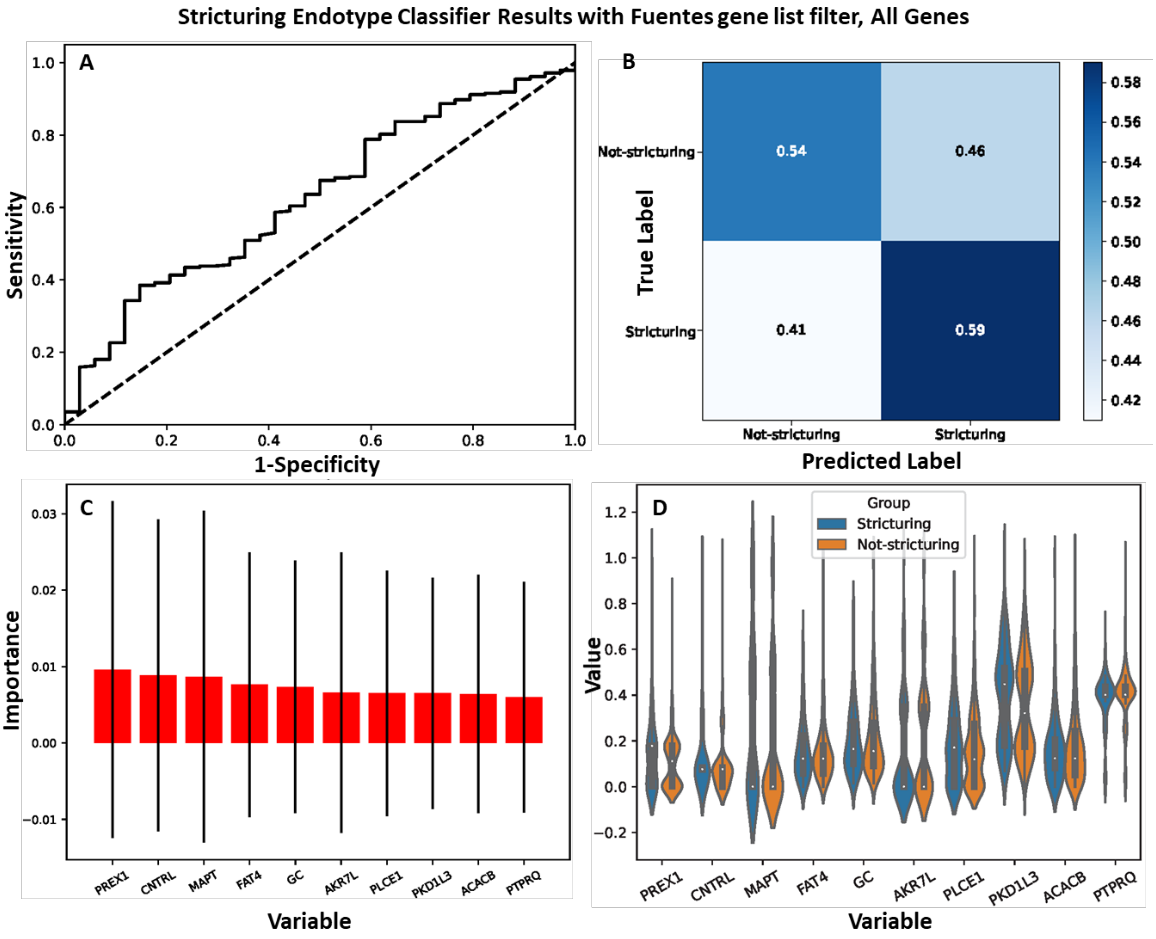


Figure 40 Stricturing endotype random forest model using the Fuentes false positive gene list as an additional filter, and the GenePy matrix with the CADD cut-off. A) Test dataset AUC; B) Normalised confusion matrix on test dataset. C) Top 10 most discriminate

genes and their relative importance for the random forest; D) Violin plots of the top 10 most discriminant genes (stricturing=blue, not-stricturing=orange).

A further examination of the contributions that different genes made to classification of CD patients by stricturing endotype by the random forest model was conducted by producing SHAP values (Figure 41A). SHAP values revealed that, of the genes visualised, a small majority showed high GenePy scores corresponding to negative SHAP values, indicating that mutations in these selected genes have the possibility of protecting against the formation of strictures (for example *MAPT*, *CNTRL* and *PREX1*). The remaining eight genes showed high GenePy scores conveying risk of stricture, including *AKR7L*, *RASAL1* and *UTP20*. There was also an extended look at the feature importances, as shown in previous figures such as Figure 40C. In Figure 41B, the feature importance value of the top 50 genes (534 genes were selected in total by feature selection, as documented in Table 23) are shown. The feature importances of all 534 genes sum to 1, and this visualises the small contributions each gene makes to classification. Unlike the subtype classifier, where *NOD2*’s feature importance was much higher that all other features, there is no stand-out gene or genes that makes a comparatively higher contribution to classification than the rest of the selected genes.

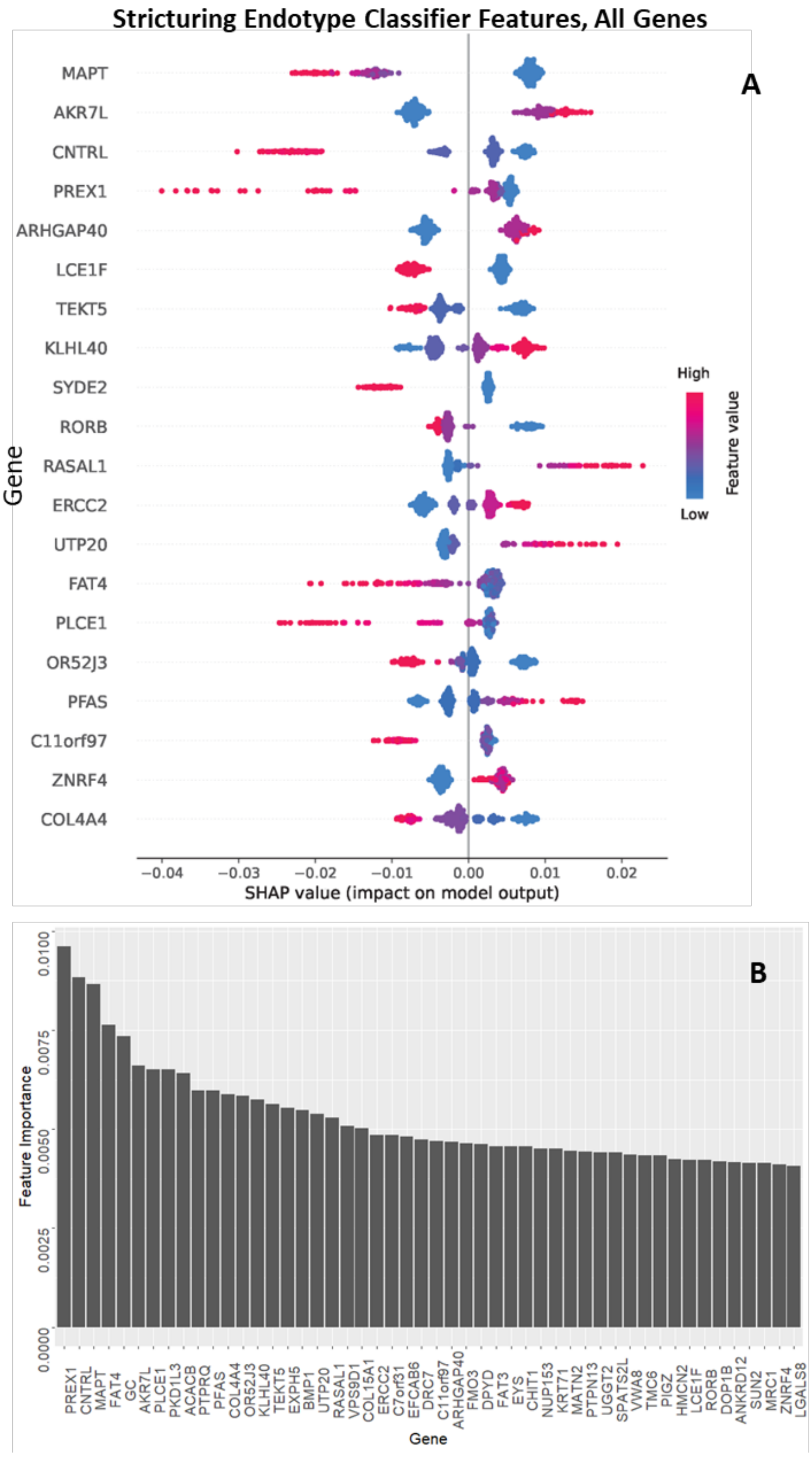


Figure 41 Further analysis of gene contributions to the best stricturing endotype classifier, using the GenePy (CADD cut-off) matrix, and the Fuentes filter. A) SHAP values for top discriminatory genes, where a high feature value is equivalent to a high GenePy score and vice versa. A positive SHAP value indicates the feature makes a contribution to the positive class, which was coding as presence of a stricture. B) Feature importance as in Figure 40C, but extended to the top 50 genes.

As performed for the subtype classifier, pathway enrichment analysis for the highest performing stricturing endotype classifier model was achieved with Enrichr [454] and the KEGG [455]2021 Human database. As the best performing classifier used all genes, the issue of a gene panel artificially enriching the pathways that corresponded to genes picked during feature selection was not present. Enrichr displayed 260 pathway terms associated with the stricturing endotype gene list. Of these, none were found to be significant after adjusting for multiple hypotheses testing (adjusted p-value < 0.05).

5.3.3 Classifying CD and UC cohorts by age of onset

In order to determine whether the underlying genomics was significantly different depending on age of onset, the RF pipeline was utilised, with the genomic data filtering which had been determined to give the best results (GenePy (CADD cut-off) matrix, Fuentes false positive gene list). This ML modelling was done for each subtype separately, such that genomic differences between CD and UC did not influence or overshadow differences between paediatric and adult onset. Here, paediatric onset IBD was defined as a receiving a diagnosis prior to 18, and all individuals receiving a diagnosis at 18 or over were defined as adult-onset IBD. As before, three gene sets were employed, 1) all available genes; 2) the autoimmune gene panel; 3) the IBD gene panel. In Table 25, the number of genes after each pre-processing filtration step are recorded for the CD age of onset classifier and the UC age of onset classifier. In Table 26 the training and testing datasets for both classifiers are recorded.

Table 25 Number of genes with GenePy scores in the GenePy (CADD cut-off) matrix at each stage of pre-processing the data prior to machine learning for age of onset classifiers. Also includes the percentage change between the genes with GenePy scores.

		Total genes with GenePy scores	**Genes after exclusion using false positive list**	**Genes with GenePy score variance**	**% Change**
CD data (n=600)	All genes	15,669	15,242	14,375	8.3%
	Autoimmune gene panel	1,598	1,586	1,486	7.0%
	IBD gene panel	499	494	468	6.2%
UC data (n=306)	All genes	15,669	15,242	13,153	16.1%
	Autoimmune gene panel	1,598	1,586	1,327	17.0%
	IBD gene panel	499	494	422	15.4%

Table 26 Training and testing dataset sizes for the CD age of onset classifier and the UC age of onset classifier

	Training Dataset		**Testing Dataset**		**Total**
	Paediatric onset	*Adult onset*	*Paediatric onset*	*Adult onset*	
CD age of onset classifier	212	212	122	54	600
	Paediatric onset	*Adult onset*	*Paediatric onset*	*Adult onset*	
UC age of onset classifier	119	119	38	30	306

The ML metrics on the testing dataset for each of the classifiers are documented in Table 27. The all genes classifier for both CD age of onset and UC age of onset classifiers achieved very high AUCs (0.92 and 0.96, respectively). These classifiers both had the same top 3 genes, *MAPT*, *APOL5*, *PRKRA*, and these genes had the highest feature importances observed throughout ML modelling in Section 5.3. There was no overlap between the top 10 genes selected by the IBD subtype classifier utilising all genes, and the CD and UC age of onset classifiers using this same gene set. The classifiers utilising the autoimmune gene panel achieved good AUCs (0.68 for CD age of onset, 0.67 for UC age of onset). There was an overlap of three genes in the top 10 for the IBD disease subtype classifier and the CD age of onset classifier: *TNC*, *DNAH12*, and *P2RX7*. In addition there was an overlap of two genes in the top 10 for the IBD disease subtype classifier and the UC age of onset classifier: *TNS1* and *DNAH12*. Six of the top 10 genes overlapped for all three classifiers (IBD subtype classifier, CD age of onset classifier, UC age of onset classifier) when utilising the IBD panel: *GC*, *DOCK8*, *GALC*, *ERAP1*, *CD6*, *NPC1*. However, where the CD age of onset classifier obtained a moderately good AUC (0.65), the UC classifier AUC performed poorly (0.44). Feature importances and violin plots of top features for the CD age of onset classifier, for all gene sets, are visualised in Figure 42. The feature importances and violin plots of top features for the CD age of onset classifier, for all gene sets, are visualised in Figure 43.

Table 27 Random forest classifier of CD age of onset and UC age of onset classifiers (P=paediatric-onset, A=adult onset). Machine learning metrics on the test set for both for the three different feature sets

CD– ALL GENES					**CD– AUTOIMMUNE PANEL GENES**					**CD– IBD PANEL GENES**				
No. Features	605				No. Features	607				No. Features	376			
	Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1
A	0.78	0.93	0.89	0.85	A	0.42	0.57	0.65		A	0.42	0.63	0.62	0.51
P	0.96	0.89	0.93	0.92	P	0.77	0.65	0.57	0.71	P	0.79	0.62	0.63	0.70.
Average	0.91	0.90	0.91	0.90	Average	0.67	0.62	0.60	0.64	Average	0.68	0.62	0.63	0.64
AUC	0.92				AUC	0.68				AUC	0.65			
Top 10 Genes	*MAPT, APOL5, PRKRA, SERPINF2, PHF1, CHIT1, MAMDC2, ZNF681, FAT4, CELA1*				Top 10 Genes	*SERPINF2, TNC, WNK1, TEK, P2RX7, PLCL1, LOXL2, TET2, ERAP1, DNAH12*				Top 10 Genes	*GC, DOCK8, PLCL1, ERAP1, TET2, GALC, NPC1, CNTRL, CD6, ITLN1*			
UC– ALL GENES					**UC – AUTOIMMUNE PANEL GENES**					**UC – IBD PANEL GENES**				
No. Features	393				No. Features	371				No. Features	309			
	Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1
A	0.97	0.93	0.97	0.95	A	0.56	0.47	0.71	0.51	A	0.33	0.33	0.47	0.33
P	0.95	0.97	0.93	0.96	P	0.63	0.71	0.47	0.67	P	0.47	0.47	0.33	0.47
Average	0.96	0.96	0.95	0.96	Average	0.60	0.60	0.57	0.60	Average	0.41	0.41	0.40	0.41
AUC	0.96				AUC	0.67				AUC	0.44			
Top 10 Genes	*MAPT, APOL5, PRKRA, CHIT1, WDR81, CYFIP1, PHF1, DNAH3, EPS8L1, PLCE1*				Top 10 Genes	*SERPINF2, BMP8A, TNS1, ADAMTS5, SALL2, NR1H3, GALC, VWF, NCOR2, DNAH12*				Top 10 Genes	*GC, DOCK8, GALC, ERAP1, CARMIL2, TET2, CD6, NPC1, PLCL1, UBASH3A*			

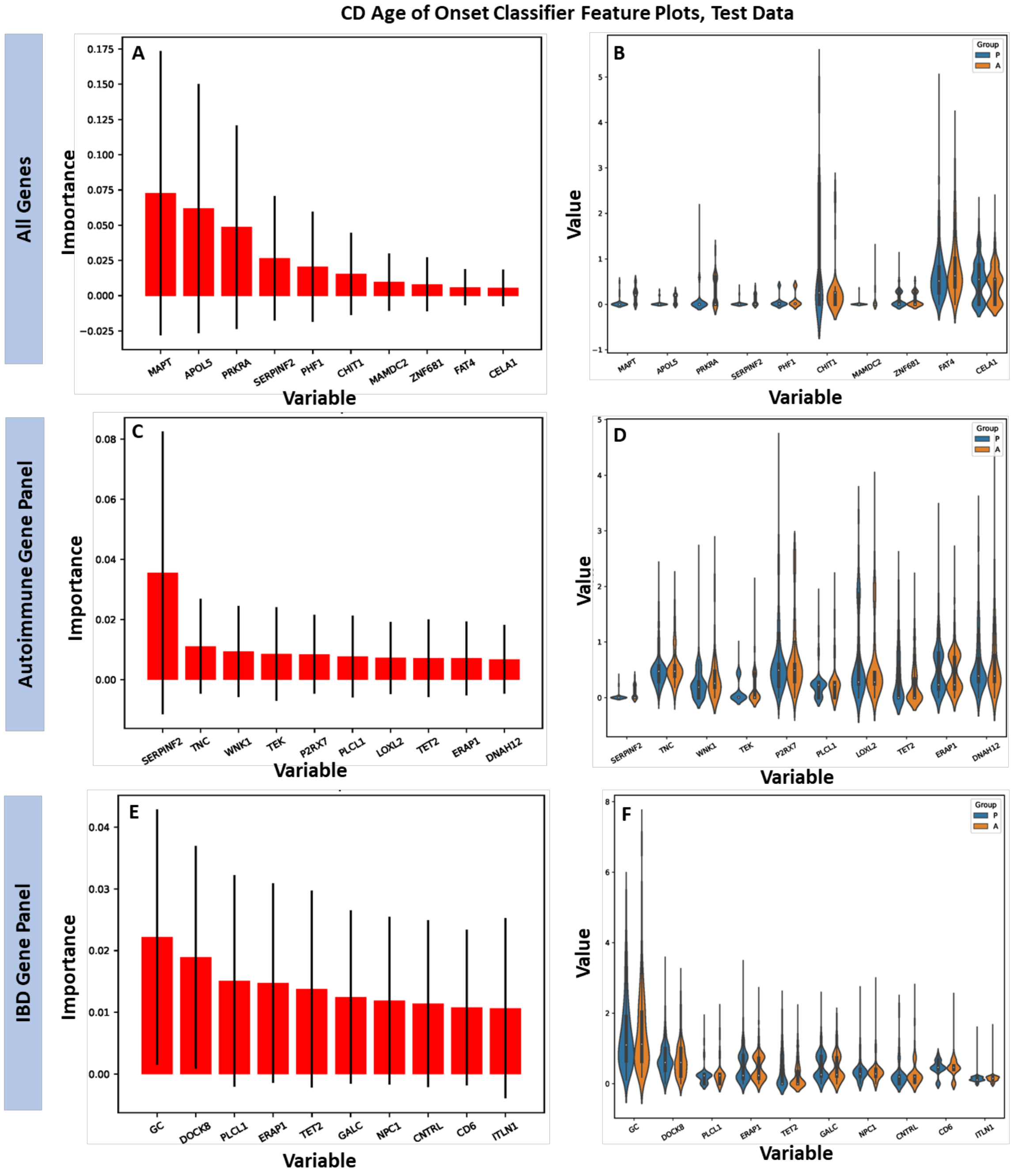


Figure 42 Feature importance and violin plots for the CD age of onset classifier for each gene set (P=paediatric onset, A= adult onset). A) Feature importances for the CD age of onset classifier which utilises all genes; B) Violin plots for the CD age of onset classifier which utilises all genes; C) Feature importances for the CD age of onset classifier which utilises the autoimmune gene panel; D) Violin plots for the CD age of onset classifier which utilises the autoimmune gene panel. E) Feature importances for the CD age of onset classifier which utilises the IBD gene panel; F) Violin plots for the CD age of onset classifier which utilises the IBD gene panel.

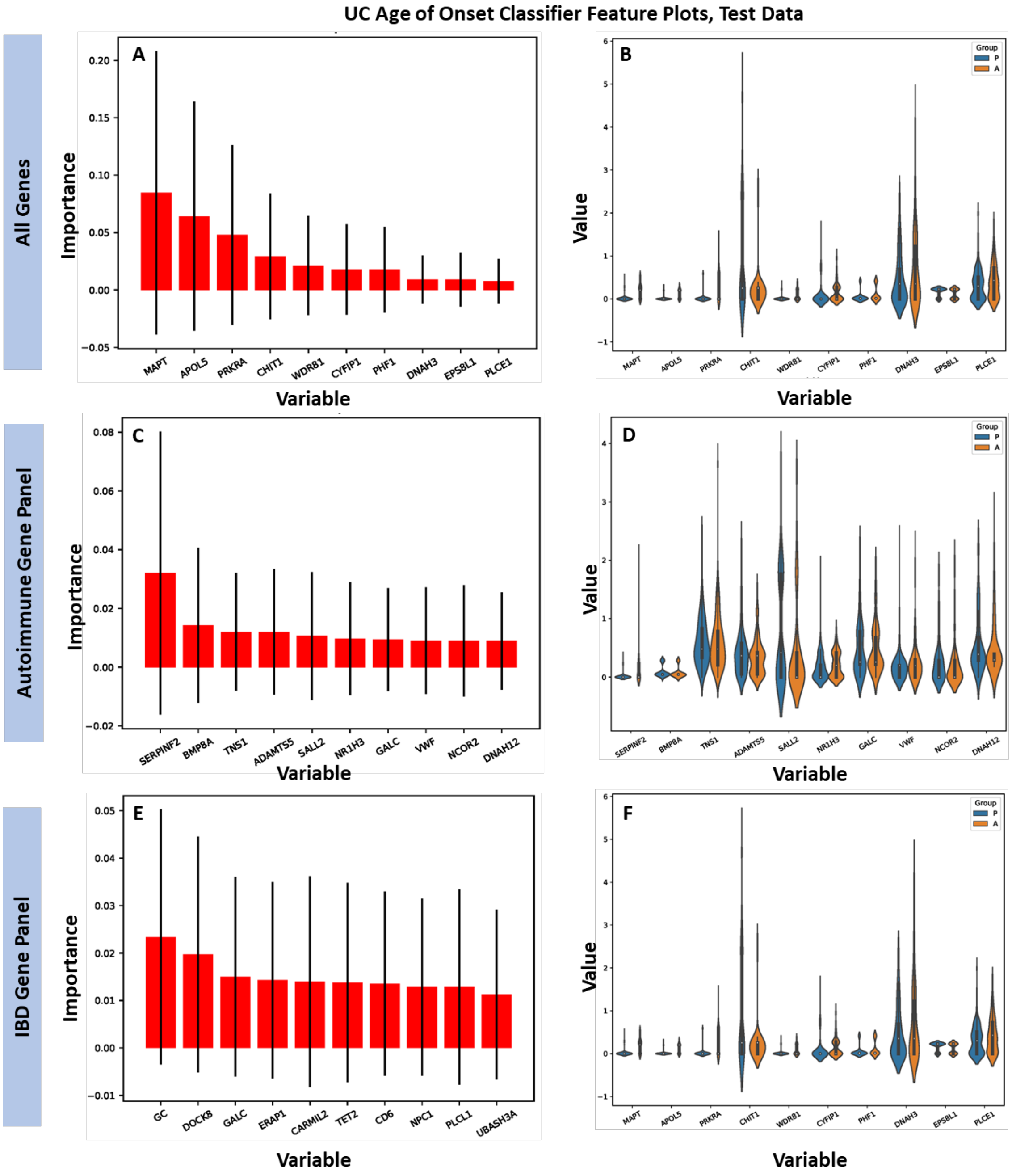


Figure 43 Feature importance and violin plots for the UC age of onset classifier for each gene set (P=paediatric onset, A= adult onset). A) Feature importances for the UC age of onset classifier which utilises all genes; B) Violin plots for the UC age of onset classifier which utilises all genes; C) Feature importances for the UC age of onset classifier which utilises the autoimmune gene panel; D) Violin plots for the UC age of onset classifier which utilises the autoimmune gene panel. E) Feature importances for the UC age of onset classifier which utilises the IBD gene panel; F) Violin plots for the UC age of onset classifier which utilises the IBD gene panel.

5.4 Discussion

Here, random forest algorithms were applied to IBD subtype determination, and classifying stricturing endotype status. In addition, through these two classification tasks, it was determined that the GenePy (CADD cut-off) matrix, and the Fuentes false positive gene lists were two beneficial genomic data processing and filtering changes. Their implementation either led to an improvement in model AUC for both tasks, or to no change, and removed genes that were known to not include pathogenic variation (for example, *TTN*). The number of variants included in the GenePy (CADD cut-off) matrix was roughly a third of the total variants available for inclusion in GenePy scores. The threshold employed here may have been too stringent. Further work could explore introducing a threshold of a Phred-scaled score ≥ 10, to see if this is a better balance between excluding trivial variants and removing variants that do contribute to a pathogenicity burden already present due to other, more damaging variants. *TTN* is also the longest gene, and this raises the question of whether gene length could affect GenePy scores, and therefore affect ML classifier results. However, as these classifiers are determining the differences between two classes, the length of the gene is constant across classes. As in, there is no comparison between the scores achieved by different genes, only the gene scores per class. Therefore, highly polymorphic genes known to be unrelated to disease were the concern, and using the Fuentes false positive gene list as a filter made excluding these genes straightforward. However, the gene remapping that had to be performed in order to utilise it did highlight the age of the original study. There is a need to re-perform analysis as performed by Fuentes Fajardo et al. [442], in light of improved high-throughput sequencing methods and bioinformatic tools, as well as availability of the GRCh38 genome build. A new list of genes that are highly likely to contain false positive pathogenic variants would be helpful for many genomic modelling approaches, and for genetic diagnostics.

In all cases of ML modelling, it is important to look at the composition of input data, as it is well known that biased inputs can affect algorithm training, and result in biased outcome predictions. One limitation of the modelling is that it only included those with European ancestry, and therefore the RF algorithm is biased towards predicting according to patterns established for that ancestral group. Restricting to one ancestry group was performed to reduce bias, to avoid genomic differences related to ancestry affecting subtype and stricturing endotype classifications. The advantage of the ML pipeline is that once a GenePy matrix has been produced a new model could be trained for groups with different genetic ancestries, given a sufficiently large cohort. Another notable bias within the clinical characteristics of the cohort was the percentage of male paediatric CD patients. It is consistently reported that rates of CD diagnosis are higher in paediatric males [14, 19], with one cohort containing 59.4% paediatric male CD patients [19]. In the IBD cohort utilised here 57% of individuals with paediatric CD were male, and when restricting to individuals who were included in the subtype model (after patient data filtering, Table 16) 58% of paediatric CD cases were male. Therefore, the male/female differences in CD paediatric diagnosis observed could be considered to be representative of patient populations observed in clinics. There is a possibility that there are some individuals within the dataset that have been diagnosed with the incorrect subtype. Where misclassification occurs, this is likely to be a CD patient misclassified as a UC patient [8]. Therefore, it is highly unlikely that the stricturing endotype classifier would be affected by subtype misclassification. When considering the subtype classifier, 26 patients that were included had minimal follow-up time of less than year. Of these, 15 patients had a UC diagnosis. In this group of 26 patients, there were no cases of infantile or VEO IBD, which can be the cases most prone to misclassification [7]. It is unlikely that subtype misclassification has biased RF algorithm predictions, but this cannot be dismissed as a possibility.

In the full IBD cohort (Table 15) 57% of those with a stricturing endotype were adult-onset CD. When considering only those included in the stricturing endotype classifier (after patient filtering by ancestry prediction and relatedness), the adult-onset stricturing endotype patients are 58% of the total. Stricturing endotype rates have been found to be similar in paediatric and adult populations after 5 years of follow up [19]. Therefore, this discrepancy is more likely due to more active and recent recruitment to the Genetics of IBD study in paediatric clinics, leading to a greater proportion of paediatric onset patients at an early stage in their disease course. Of those included in the stricturing endotype classifier, 41% of paediatric-onset CD have less than 5 years follow-up, in comparison to 6% of adult-onset CD patients.

For the disease subtype classification model, the best result was obtained using the autoimmune gene panel (AUC 0.67). Regardless of the gene panel utilised in ML modelling, *NOD2* was always ranked as the most important gene. This was particularly impressive for the ML model that used all genes. This random forest had no prior gene filtering based on biological knowledge, and still singled out *NOD2* as an important genetic discriminant of the two IBD subtypes. The different AUCs achieved through the use of each gene panel highlight a difficult balance when using biological knowledge as feature selection. Including all possible genes creates problems for ML modelling, as it increases the dimensionality of the dataset with little to no advantage; many genes that are highly unlikely to be associated with IBD pathogenesis are included, with the upside that a few undiscovered genes will also be included. On the other hand, the panel which solely focuses on genes implicated in IBD does not perform as well as the autoimmune gene panel. Clearly, a gene panel that is too restrictive risks missing important genes for classification, and also important gene-gene linear or non-linear interactions.

Aside from *NOD2* being selected as the most discriminant gene in random forest modelling, another gene present in the NOD-signalling pathway, *P2RX7*, was also in the top 10 important genes. This is a key innate immune pathway highly implicated in CD aetiology [102, 458]. Other genes of note in the top 10 most discriminant genes are *WDFY4*, *TNC* and *NFATC1*. Interestingly, *WDFY4* has previously been reported as a gene associated with systemic lupus erythematosus, and not CD or UC in a GWAS meta-analysis of risk loci associated with autoimmune diseases [459]. However, in the violin plots, there is a clear tail to the GenePy score distribution in the CD group. This suggest some rare variation that has a high CADD score is present in a subset of these patients, potentially rare enough to not be detected in GWAS. *WDFY4* is thought to be involved in autophagy [460]. When levels of the glycoprotein product of *TNC*, Tenascin-C, were measured in IBD patients, they were found to have elevated levels in comparison to controls [461]. The gene's glycoprotein is involved in arresting T-cell activation and intestinal barrier function. More recently *TNC* was associated with IBD during a GWAS performed with an African American IBD cohort [462]. It is possible that variation in this gene is present in individuals with European ancestry, but it is rarer in that cohort than in those with African ancestry, again meaning GWAS on a European cohort would not detect this. Finally, *NFATC1* plays a role in T-cell activation, in particular in the induction of *IL-2* and *IL-4* [463]. It was also present in the top 10 important genes in the classifier that used the IBD gene panel.

Further investigation into the contribution of genes to subtype classification involved producing SHAP values and visualising the feature importances of an increased number of genes. A trend which emerged in the SHAP values was that a higher GenePy score (feature value) was associated with a positive SHAP value, meaning those values contributed to discriminating individuals as CD. For a few genes, such as *IL31RA*, *NRP1* and *LRP1*, high GenePy scores were associated with UC classification, but far more common was a low GenePy score (minimal-to-no variation in that gene) contributing to discriminating UC cases. This is reflective of current biological knowledge, whereby the percentage of genetic heritability that has been accounted for is higher for CD than UC [68]. Through SHAP values and feature importance plots, *NOD2* cements itself as the strongest predictor, with very clear delineation between positive and negative SHAP values. Plotting the feature importances of the top 50 genes emphasises, after *NOD2*, how small a contribution each gene makes to discriminating between each subtype class. This is one reason why a network analysis approach such as STRING [464] was considered inappropriate. This visualisation of SHAP values and feature importances does not reveal any subset within the 739 features selected that were more important to classification. Therefore, a cutoff cannot be established for a gene subset, and a very large gene network would be minimally informative. Further, the autoimmune gene panel is already enriched for genes that are related to similar mechanisms, and finding that these genes interact would be expected. This is a disadvantage in using panels, as although it decreases the dimensionality of the input data, the gene choices are predicated on biological knowledge, and therefore these genes will often be part of related pathways.

A traditional pathway enrichment analysis approach was revealed to be inappropriate for the data, as the initial autoimmune gene panel was already enriched for key immune pathways, hence artificially enriching the pathways of the genes selected. After this, the pathways which were no longer significant were deemed to be more informative. An interesting result from this pathway enrichment analysis approach was the exclusion of pathways associated with type I diabetes mellitus, and autoimmune thyroid disease. It has been established that many autoimmune diseases co-occur [168-170], and this is thought to be due to underlying immune dysfunction manifesting as more than one autoimmune disease. It is therefore interesting, and potentially useful for exclusion of genes from further investigation. Whether autoimmune thyroid diseases and type 1 diabetes co-occur with IBD at lower rates than other autoimmune diseases could also be a subject for further investigation. However, when performing further investigations into gene associations it is important to remember that the relationship between the genes chosen (or not chosen) during feature selection and the chosen modelling outcome is only as strong as the testing AUC achieved. This is because the testing AUC is representative of the generalisability, and reliability of the results. The testing AUC achieved, in combination with small SHAP values and feature importances for genes aside from NOD2, combine to give very limited confidence in any predictions that could be made with related genes and pathways.

For the stricturing endotype classifier, the best performance was achieved with the random forest that utilised all available genes (AUC 0.63). It was surprising to see the mediocre performance of the more targeted autoimmune and IBD panels, where the AUCs were only slightly better than random. Many factors could be at play here for this classifier to produce a worse performance than the disease subtype classifier. Firstly, while the patient groups for CD and UC are approximately as imbalanced as the stricturing endotype groups (1:2 UC:CD, 1:2.5 stricturing:not-stricturing), the stricturing classifier is performed on only CD patients, and as such this training data is approximately half the size of the disease subtype classifier's training data. Additionally, the gene panels used here were geared towards IBD, and there was no stricturing endotype specific panel. As established for the disease subtype classifier, the gene panel does play an important role. Therefore, a more bespoke panel may produce better results. Further, the not-stricturing clinical group is not as certain as the stricturing group. Patients within the not-stricturing group may stricture in the future. The ML model may have predicted some individuals as stricturing when they are in the not-stricturing group, and thus this is an incorrect classification currently, but the algorithm's prediction may be proved correct in the future. It is therefore reassuring that the ML model was better at identifying individuals in the stricturing class, than the not-stricturing class (sensitivity 0.59 vs 0.54).

Of the genes selected as the most discriminant by the stricturing endotype classifier utilising all genes, *PREX1*, *GC*, and *PLCE1* are of note as having connections to inflammation and the immune system, or IBD pathogenesis. *PREX1* is associated with innate and adaptive immunity and is a potential target of microRNAs that were found to be overexpressed in CD and UC patients [465]. *GC* is also known as the Vitamin D Binding Protein (*VDBP*). Studies have shown lower levels of VDBP in paediatric IBD patients than healthy controls [466], and higher VDBP concentrations were associated with an increased risk of disease flare in adult CD patients [467]. SNPs in *PLCE1* have been reported as associated with colorectal cancer [468], and the gene is associated with MAPK signalling, which can initiate inflammatory processes [469].

When observing the SHAP values produced from the stricturing endotype classifier utilising all genes, there are some genes where a higher GenePy score (feature value) have a negative SHAP value, therefore contributing to a not-stricturing classification and implying variation in those genes is protective. For other genes, the more expected relationship between stricturing endotype (positive SHAP value) and high GenePy scores is observed. Of note is that for *CNTRL*, *TEKT5*, and *PFAS*, there are data points for low or lower GenePy scores (feature values), which correspond to both positive and negative SHAP values. This arguably indicates that this model isn't a strong discriminator of the two classes in comparison to the subtype classifier. The extended feature importance plot of the top 50 genes emphasises the small contribution each of the 534 features makes to the discriminating stricturing and not-stricturing statuses. Pathway analysis did not reveal any significantly enriched pathways. As no panel was used, this was an agnostic approach, where the feature selection had the potential to choose genes that belong to pathways previously not associated with the stricturing endotype. No significantly enriched pathways suggests that the highly dimensional nature of the dataset, where the number of genes (features) greatly exceeds the number of individuals (samples) has led to challenges in feature selection.

An analysis of the potential differences in the underlying genomics of IBD patients depending on age of onset was performed using the random forest ML pipeline. This problem was transformed into a binary classifier of paediatric (< 18 years of age at diagnosis) and adult onset (18 years and over at diagnosis). When using all genes, the CD age of onset classifier, and the UC age of onset classifier attained very high testing AUCs of above 0.9. The top three genes were the same for both classifiers: *MAPT*, a gene encoding microtubules that is differentially expressed in the nervous system [470]; *APOL5*, a gene encoding a cytoplasm protein that may affect lipid movement [471]; and *PRKRA*, a protein kinase that mediates the effects of interferon in response to viral infection [472]. None of the top 10 genes in these age of onset classifiers were selected in the IBD subtype classifier. The combined feature importances of the aforementioned genes sum to 0.18 for the CD age of onset classifier, and 0.20 for the UC age of onset classifier. This goes against the trend observed in IBD subtype classifiers and stricturing endotype classifiers of small feature importances attributed to each gene. Given this different trend in gene importances; that genes are not related to any pathways or functions known to contribute to IBD development; and that there are no top genes in common between the age of onset classifiers and the IBD subtype classifier, this classifier was thought to be unreliable, with the high AUC observed potentially the result of data artifacts. Regardless of cause, these differences in genes do not appear to impact the IBD subtype classifier. For the autoimmune gene panel classifiers there were 2 overlapping genes (*TNS1* and *DNAH12)* in the top 10 of the CD and UC age of onset classifiers, and the IBD subtype classifiers. For the IBD gene classifiers 6 genes overlapped between the three classifiers (*GC*, *DOCK8*, *GALC*, *ERAP1*, *CD6*, and *NPC1*), although the UC age of onset classifier was unable to discriminate paediatric and adult IBD (AUC 0.44). Observing the violin plots for all classifiers produced leads to the possible hypothesis that rare variation, (long tails on the GenePy score distributions) drive these classifications. This is similar to what was observed for the IBD subtype classifier. In general, genes appear to have similar distributions with more extreme scores appearing in both paediatric and adult onset groups, depending on the gene. These more extreme scores may then go on to drive the IBD subtype classifier, regardless of which age of onset group is more likely to have these higher GenePy scores.

During identification of adult IBD patient's stricturing endotype status, the number of clinical records required to review was reduced by using keyword flags. However, over 1,000 records were still reviewed manually. The process was time-consuming, and error-prone. The record fields searched were free-text fields, and as such some records could be more ambiguous than others. This meant additional checks were required by a clinician in order to verify stricturing endotype status where it was not clear. This adds to the time required to gather this patient information.

Another challenge of collecting this data, is its potential to change quickly. Unlike a subtype diagnosis, CD patients are monitored regularly, and have the potential to develop stricturing behaviour at any time. The stricturing endotype data collection could therefore become out of data very quickly. There is a requirement to streamline records for patients in order that complications like stricturing can be easily identified. Other endotypes could also be investigated if clinical records were more automated. An example of this is the fistulating endotype, where a tunnel can form connecting one portion of the bowel to another section, or to the outside of the body. ML models for these specific prognostic questions cannot be generated unless there is a quick, reliable way to gather this clinical data.

The random forest ML performed here produced a good AUC for the disease subtype classifier, and a modest AUC for the stricturing endotype classifier. These results present the real potential of utilising genomic data and ML for IBD. However, there are additional measures that could be implemented to potentially improve the algorithm's performance. Aside from including more trees (estimators = 10,000) in the random forest, all other hyperparameters were set to the default for random forest. Optimisation of these hyperparameters, which dictate the rules of the random forest algorithm, could lead to improved classification results. It is unfortunate that the stricturing endotype classifier did not perform as well as the disease subtype classifier. Being able to predict a patient's disease course, and whether they are susceptible to the development of specific endotypes and complications would have more impact on patient management and their quality of life than the prediction of their disease subtype. As discussed above, the use of another gene panel for the stricturing endotype classifier could lead to better random forest performance, which would be a step towards personalised medicine.

Chapter 6 Optimisation of machine learning for inflammatory bowel disease subtypes and the Crohn’s disease stricturing endotype

Chapter summary – this chapter focusses on optimising the modelling performed in Chapter 5, for IBD subtypes, and the CD stricturing endotype. Two hyperparameter tuning methods were investigated, and their results compared for the different clinical tasks. In addition, for the stricturing endotype additional optimisation of the genomic and patient input data was performed.

Chapter contributions – Whole exome sequencing data was joint-called by Guo Cheng, with all subsequent processing, and transformation into GenePy scores performed by Imogen Stafford. The IBD gene panel was generated through literature searches performed by Guo Cheng and James Ashton. Stricturing gene panel literature search was performed by Imogen Stafford. Clinical stricturing status and follow-up were assessed by Imogen Stafford and James Ashton. Modelling and optimisation were performed by Imogen Stafford, with guidance from Mahesan Niranjan.

Supplementary files can be found at https://doi.org/10.5258/SOTON/D2655.

6.1 Introduction

6.1.1 Hyperparameter tuning

When optimising a machine learning (ML) model to obtain the best results, there are two aspects of the algorithm to consider: the parameters and the hyperparameters. Parameter optimisation occurs during the training of a model. These parameters will be defined in relation to the prediction task and the input data. The parameters of the model will determine how the data is classified. The input data and prediction task are also used for hyperparameter tuning, but in this case what is determined are the constraints on how the algorithm can operate to subsequently classify the data (descriptions of random forest hyperparameters are included in Table 28). For example, if data is very noisy, then to classify the data with a random forest it could be beneficial to set the hyperparameters such that the model does not become too complex. This could be achieved by reducing the maximum depth of each tree, so that there are few data splits per tree, or the number of samples required to make a tree split could be increased to prevent many splits resulting in end nodes with one sample in each. However, it is difficult to know intuitively what the best hyperparameters could be for the best classification result.

There are several approaches that can be taken in order to obtain the optimal hyperparameters for a model. An exhaustive grid search will work through all possible combinations of hyperparameters, training a model for each in order to determine the optimal set. While thorough, this approach can be very time intensive. The random search method will select a hyperparameter combination at random and train a model with these hyperparameters. A second set of hyperparameters will be selected, model trained, and these results are compared. The algorithm then retains the better hyperparameter combination. This will continue for the number of hyperparameter trials selected by the user. In both these methods, there are no assumptions about the potential best hyperparameter, and every trial of a hyperparameter combination is independent. A downside of the random search method is that, while the algorithm tries to minimise the cost function, i.e. create the most accurate model, it can find, and subsequently be trapped, in a local minima. Then, the random search will select a combination of hyperparameters that will give a good machine learning model result, but not the best that could have been achieved, if the search had found the global minimum. An alternative to these is a Bayesian approach. A Bayesian optimisation search will use the performance of the previous hyperparameter combination to inform the choice of the next hyperparameter combination. Similar to random search, the number of hyperparameter combinations trialled is selected prior to tuning. In addition, the Bayesian approach samples points across the cost function to identify possible minima. It is this combination of trialling all possible minima, and prior knowledge, that means this approach is almost certain to arrive at the global minimum and give the optimal hyperparameter combination for the creation of the best model.

Table 28 Description of the function of the random forest hyperparameters optimised in this chapter.

Hyperparameter Name	Python Variable Name	Definition
Number of estimators	n_estimators	The random forest model is an ensemble classifier that outputs its classification results based on many decision trees. This parameter determines the number of trees generated during the modelling.
Maximum Tree Depth	max_depth	The maximum number of times that the data is split (a decision) in each tree.
Minimum samples per split	min_samples_split	The minimum number of samples required at a tree node to split the data.
Minimum samples per leaf	min_samples_leaf	The minimum number of samples required in a leaf node (a node which classifies the data present at that node into a category).
Maximum Features per split	max_features	The maximum number of features in the dataset to consider for each tree split.

6.1.2 Nested Cross-validation

The principle of cross-validation was introduced in Section 1.3.5, and for hyperparameter optimisation using cross-validation gives a more generalised and robust estimate of the performance of each hyperparameter combination. However, a potential issue when using a simple cross-validation scheme to optimise a model’s hyperparameters, is that during this process both the parameters determined from the data, and the hyperparameters of the model will be decided upon. It is therefore possible that information leaks through from the hyperparameter tuning to decisions regarding model parameters. As a result, the estimated performance of the tuned model can be inflated. One approach to minimising information leakage is to use a nested cross-validation scheme illustrated in Figure 44. An inner cross-validation is performed using the training data of the outer cross-validation. Therefore, a separate combination of data is used to determine the optimum hyperparameters. The ML algorithm, and its parameters, is subsequently trained and tested on the outer cross-validation data.

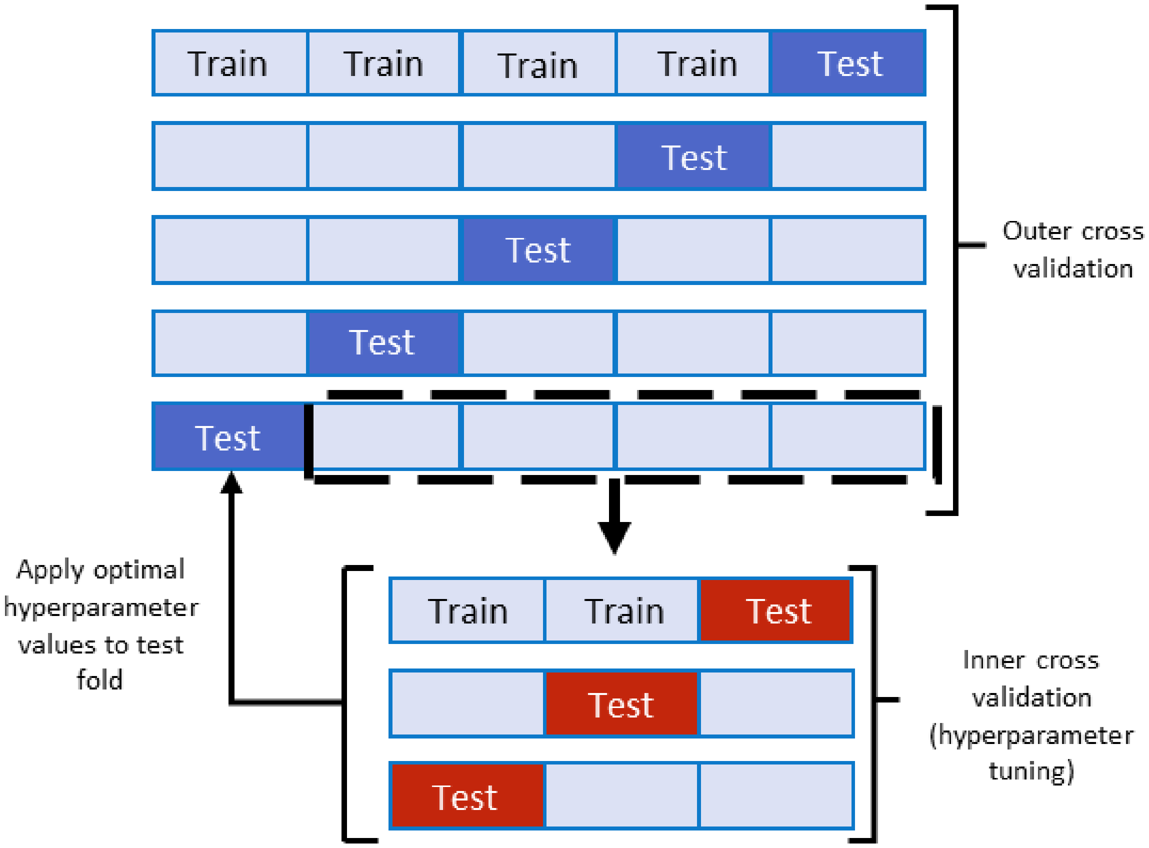

Train
Train
Train
Train
Test
Test
Test
Test
Test
Outer cross validation
Apply optimal hyperparameter values to test fold
Train
Train
Test
Test
Test
Inner cross validation (hyperparameter tuning)

Figure 44 Example of a nested cross-validation scheme that uses 5-fold outer cross-validation and 3-fold inner cross-validation.

This chapter begins with the machine learning pipeline established in Chapter 5, applied to two classification problems: 1) IBD subtypes; and 2) CD stricturing endotype. For both clinical tasks, hyperparameter tuning is performed in three different ways: I) individually, to observe each hyperparameter's behaviour, II) using the Grid Search method, and III) using Bayes optimisation. By selecting the best combination of hyperparameters, the machine learning models become more tailored to the idiosyncrasies of the genomic data, and by extension the individual clinical classification problem. For the stricturing endotype classifier, multiple gene panels were trialled to try and arrive at an optimal gene set that best classified the patients. In addition, a filtered patient dataset was used in some modelling to observe if requiring a specific number of years of clinical follow-up in the 'not stricturing' group would enable a model to better distinguish between the stricturing and not-stricturing groups.

6.2 Methods

In Chapter 5, two main GenePy score processing steps were tested: a CADD Phred filter that only included likely pathogenic variants were included in GenePy scores, and the Fuentes false positive gene list [442] so that genes with a high pathogenicity burden, but have been identified as not disease causal, would not be included in modelling. These measures were intended to reduce noise in the dataset, giving the random forest modelling clearer genomic signals to detect. As these steps were successful, they were implemented on the GenePy score matrix before hyperparameter tuning. Other standard pre-processing of the GenePy score matrix and patient data was performed as in Section 5.2.4 and Section 5.2.5. To begin optimisation of the modelling, the gene panel representing the genomic data input was determined. For the disease subtype classification task, this was the autoimmune gene panel, as this data input gave the best random forest modelling results in Section 5.3.2. For the stricturing endotype classifier, during previous modelling no one gene set was determined as the optimal one. Therefore, as part of model optimisation discussed herein, several gene panels were evaluated:

I) All genes: the genes that GenePy scores could be generated on.

II) Autoimmune panel: the HTG EdgeSeq panel

III) IBD panel: the in-house IBD panel that includes genes associated with IBD-like monogenic illness, and genes identified through assessment and analysis of IBD GWAS (unpublished data)

IV) Extended NOD signalling pathway: comprises genes included in the KEGG:hsa04621 and REACT:R-HSA-168638 pathways.

V) Stricturing panel (inclusive): includes genes from a literature search of genes associated with the stricturing endotype, and all genes from (II) and (III).

VI) Stricturing panel (exclusive): only includes genes from the stricturing endotype literature search and does not include genes from (II) and (III) unless they were identified in the literature search.

As stated above, a literature search was performed to collate a comprehensive list of genes implicated in the development of a stricturing endotype. A Boolean literature search was performed in PubMed. There was a low threshold for including genes in the stricturing panel, to try and assemble an inclusive list. The search was as follows: (Crohn’s disease OR Crohn disease) AND (stricture OR stricturing OR fibrotic OR fibrosis) AND (gene OR genetic). Papers from 2016 to the present (search performed 20th September 2021) were assessed. Genes were included in the stricturing panel if they were implicated as a causal or a protective gene.

In addition, the patient data included in the stricturing endotype model was also considered. The follow up period in the IBD cohort is highly varied. The study that recruits patients to this cohort is ongoing, so patients recruited in recent months and years will have little or no follow-up data, and in most cases not enough time will have passed for these patients to develop strictures. This means there may be some patients that are currently classified as not-stricturing who may develop a stricturing endotype in the future. Therefore, two patient datasets were used: one which included all patients, and one that set a follow-up threshold in the not-stricturing group to exclude patients where their future stricturing endotype status is uncertain. The duration of a patient's follow-up time was determined by the date of most recent clinical contact, which was defined as either pathology results, an outpatient appointment or an admission to hospital.

Feature selection for reduction of the number of genes used in modelling was performed with a support vector classifier as in the previous chapter. Next, three different hyperparameter tuning processes were completed:

- **Individual hyperparameter tuning:** five random forest classifier hyperparameters (max_features, n_estimators, min_samples_split, min_samples_leaf, and max_features) were tuned utilising the GridSearchCV algorithm, contained within the Python (v.3.7) package scikit-learn [446]. A non-nested cross-validation approach, with 7 folds was used, as measuring the generalisability of these models was not required.
- **Grid Search hyperparameter tuning:** GridSearchCV was used to tune all five hyperparameters simultaneously, given a limited number of values that each hyperparameter could take. This was done within a nested scheme, with 7-folds in the outer cross-validation, and 5-folds in the inner cross-validation.
- **Bayes Search hyperparameter tuning:** BayesSearchCV from the Python (v.3.7) package scikit-optimize was used to perform the tuning of the five hyperparameters simultaneously. The range of values that each hyperparameter could take was wider than for the Grid Search, as the number of iterations of hyperparameter tuning is less that the former, exhaustive method. This was done within a nested scheme, with 7-folds in the outer cross-validation, and 5-folds in the inner cross-validation.

For each hyperparameter tuning process, the model performance was assessed using balanced accuracy. The aim with **individual hyperparameter tuning** was to observe how each hyperparameter could impact modelling, and to what extent (i.e. the change in balanced accuracy). For the **Grid Search** and **Bayes Search** tuning, the chosen hyperparameter combination was used in the downstream modelling in order to understand how this changed the characteristics of the model, and its performance. The hyperparameter tuning described here was performed separately for both the disease subtype and stricturing endotype classifiers.

The set of tuned hyperparameters from the Grid Search and Bayes Search was then applied to the whole training set to get a final random forest model, and this model was applied to the test set, which had not been used for any tuning or training. This resulted in the generation and assessment of three random forest models: an untuned model, a model tuned using GridSearchCV, and a model tuned using BayesSearchCV. The random forest model test set performance was assessed as before using the area under the curve, as well as other output metrics (precision, sensitivity, specificity and F1 score). Genes that contributed to the model were analysed. SHAP values [453] were produced for the disease subtype and stricturing endotype classifiers that had been hyperparameter tuned, as in Section 5.2.7. Pathway analysis with Enrichr [454] was performed for a hyperparameter tuned stricturing endotype classifier, as in Section 5.2.7. The full pipeline for these methods is illustrated in Figure 45 (see Supplementary Files for machine learning scripts).

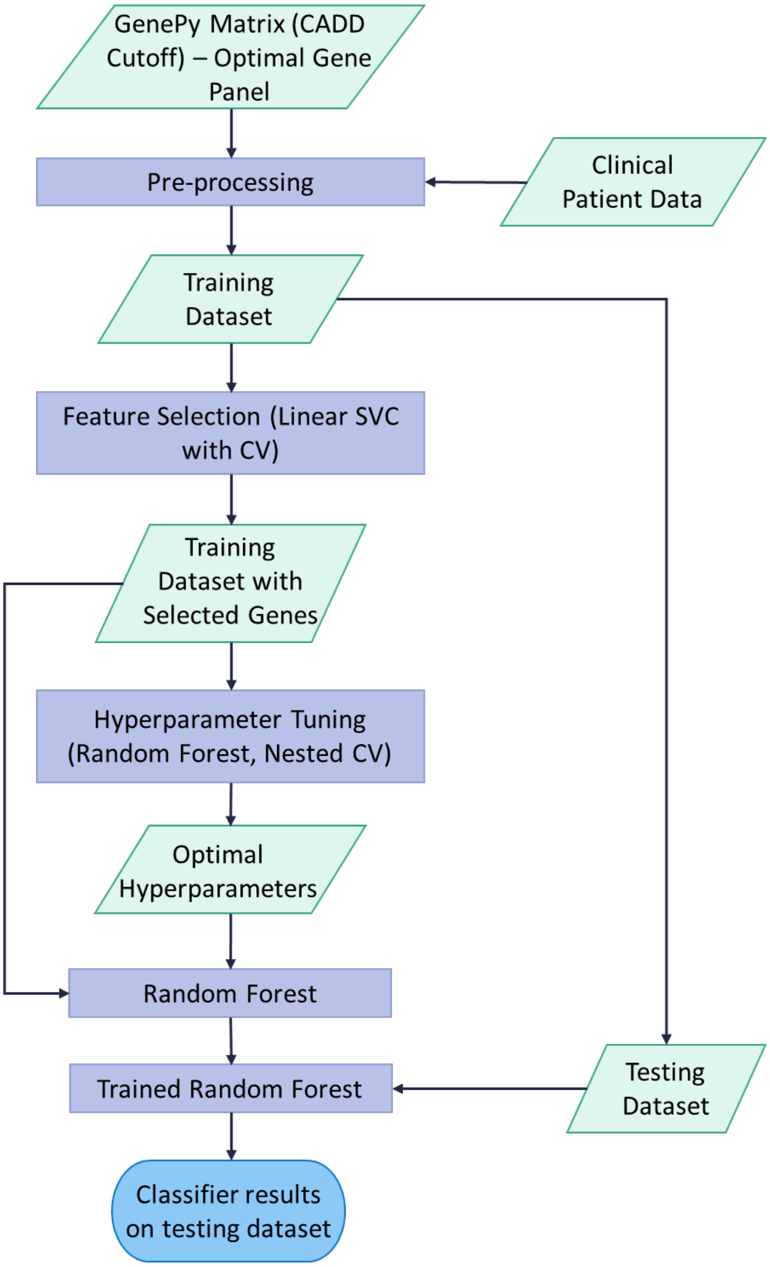

GenePy Matrix (CADD Cutoff) – Optimal Gene Panel
Pre-processing
Clinical Patient Data
Training Dataset
Feature Selection (Linear SVC with CV)
Training Dataset with Selected Genes
Hyperparameter Tuning (Random Forest, Nested CV)
Optimal Hyperparameters
Random Forest
Trained Random Forest
Testing Dataset
Classifier results on testing dataset

Figure 45 Machine learning pipeline with the addition of hyperparameter tuning. The hyperparameter optimisation step is always performed using random forest as the base algorithm, with nested cross-validation, but the method of tuning is either Grid Search or Bayesian Optimisation, depending on the approach being trialled. This pipeline was performed for both the subtype classifier and the stricturing endotype classifier.

6.3 Results

6.3.1 Disease subtype classifier

This classifier includes CD and UC patient data. After pre-processing using ancestry and relatedness information (full details in Section 5.2.5), 600 CD, and 306 UC patients are included in the modelling. The clinical characteristics of these individuals are as detailed in Section 5.3. The training dataset consists of 244 CD, and 244 UC patients (80% of patients, according to the minority class), and the testing dataset includes 356 CD, and 62 UC patients.

6.3.1.1 Hyperparameter tuning

In Table 29, the values that were trialled for each hyperparameter in each tuning method (**individual hyperparameter tuning**, **Grid Search**, **Bayes Search**) are given. These are given in a list format, as unlike other classifiers, such as support vector machine, where the hyperparameter tuning space is continuous, each hyperparameter for random forest expects an integer as input. Due to the computational and time costs of Grid Search, fewer hyperparameters values are included for this tuning process.

Table 29 Hyperparameter default values, and values tested during individual, grid search and Bayes hyperparameter tuning for the disease subtype classifier

Hyperparameter Name	**Default**	**Values Tested – Individual Hyperparameter Tuning**	**Values Tested – Grid Search**	**Values Tested - Bayes**
n_estimators	100	100, 250, 500, 750, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000	500, 750, 1000, 5000	100, 250, 500, 750, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000
max_depth	None	1 - 30, None	5, 10, 20, 30, None	1 - 30, None
min_samples_split	2	2, 3, 4, 5, 6, 7 ,8, 9, 10	2, 3, 4, 5	2, 3, 4, 5, 6
min_samples_leaf	1	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1, 2, 3, 4, 5	1, 2, 3, 4, 5, 6
max_features	sqrt	sqrt(n_features), log2(n_features) None	sqrt(n_features), log2(n_features) None	sqrt(n_features), log2(n_features) None

Each hyperparameter was tuned individually to gauge the potential impact each one could have on the performance of the random forest model (Figure 46). Overall, changing individual hyperparameter values resulted in only small changes to the cross-validated balanced accuracy achieved by the models. Minimum and maximum balanced accuracies attained by specific hyperparameter values were often within 0.1 of each other. The balanced accuracy range was particularly small for the minimum samples per leaf, and maximum depth hyperparameters. Hyperparameter tuning for the number of estimators had the expected trend of the balanced accuracy increasing sharply as the number of estimators increased, before reaching a plateau (Figure 46B). The minimum samples per leaf hyperparameter also exhibited a trend towards poorer model performance as the number of samples required at a leaf increased (Figure 46D). In contrast, the minimum samples per split hyperparameter did not have a clear trend that indicated an optimal value, or range of values (Figure 46A). The most successful hyperparameter, by CV balanced accuracy, was the maximum feature number. Setting this to “None” achieved a balanced accuracy of 0.6 (Figure 46C).

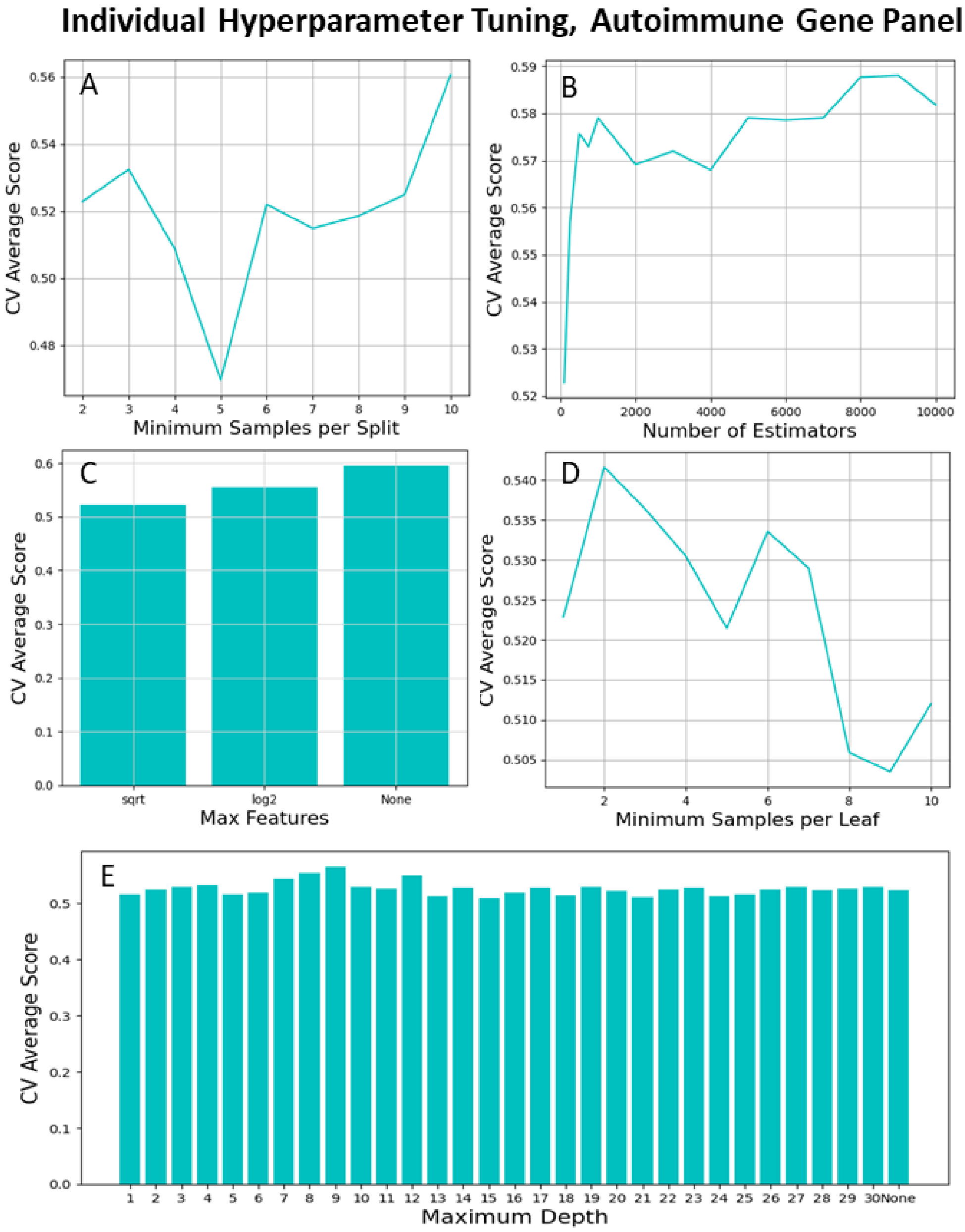


Figure 46 Average balanced accuracy score across the 7 cross-validation (CV) folds for individually tuned hyperparameters (disease subtype classifier).

Using an exhaustive Grid Search, a set of values for each hyperparameter was tuned together with nested cross-validation. Chosen hyperparameters for each inner cross-validation fold and the corresponding average balanced accuracy across five folds, and the balanced accuracy in each of the outer cross-validation folds can be viewed in Table 30. The average balanced accuracy across all 7 outer folds was 0.489 (standard deviation 0.087). There was variation in the hyperparameters chosen, particularly for the minimum number of samples per leaf hyperparameter, and the number of estimators hyperparameter. The optimal hyperparameters were chosen according to the balanced accuracy of the corresponding outer cross-validation fold. Here, the best outer fold had a balanced accuracy of 0.643 (fold 5), and the following hyperparameters were selected: maximum depth = 5, maximum features = none, minimum samples per leaf = 1, minimum samples per split = 5, and number of estimators = 1000.

Table 30 Hyperparameters selected by nested grid search and corresponding balanced accuracy in inner and outer cross-validation folds for the disease subtype classifier.

Fold	**Outer Fold Balanced Accuracy**	**Inner CV Balanced Accuracy**	**Optimal Hyperparameters**				
			Max Depth	**Max Features**	**Minimum Samples per Leaf**	**Minimum Samples per Split**	**Number of Estimators**
1	0.500	0.533	5	log2	5	2	1000
2	0.452	0.706	5	None	5	2	5000
3	0.500	0.500	5	log2	4	2	500
4	0.500	0.560	10	log2	1	5	1000
5	0.643	0.525	5	None	1	5	1000
6	0.325	0.639	5	None	1	2	750
7	0.500	0.562	5	None	4	2	750

Next, the Bayes Search method with 60 iterations was used to optimise the hyperparameters. The same nested cross-validation scheme of 5-folds in the inner cross-validation, and 7-folds in the outer cross-validation was used. The results of this optimisation, with the best inner cross-validation balanced accuracy and corresponding hyperparameters, and the balanced accuracy when applying these hyperparameters to the outer fold test set, can be viewed in Table 31. All hyperparameters, except the maximum number of features, varied widely across the different outer folds. The average balanced accuracy across the outer folds was 0.569 (standard deviation 0.034). The best balanced accuracy in the outer fold was 0.610 in fold 5. Therefore the chosen hyperparameter values from Bayes optimisation were: maximum depth = 27, maximum number of features = None, minimum samples per leaf = 1, minimum samples per split = 4, and 250 estimators.

Table 31 Bayes search nested CV results for the disease subtype classifier, with balanced accuracy in inner and outer folds for each fold's selected hyperparameter combination.

Fold	**Outer Fold Balanced Accuracy**	**Inner CV Balanced Accuracy**	**Optimal Hyperparameters**				
			Max Depth	**Max Features**	**Minimum Samples per Leaf**	**Minimum Samples per Split**	**Number of Estimators**
1	0.538	0.603	6	None	5	6	250
2	0.523	0.628	28	None	4	3	8000
3	0.548	0.611	22	None	2	3	750
4	0.585	0.578	13	None	3	2	250
5	0.610	0.604	27	None	1	4	250
6	0.559	0.624	24	None	6	6	100
7	0.619	0.571	11	Sqrt	1	5	5000

6.3.1.2 Application of optimal hyperparameters to random forest modelling

After the optimal hyperparameters were selected by Grid Search and Bayes Search methods, the random forest was trained with these hyperparameters on the whole training set. A comparison of untuned and tuned random forest model results on the test set is recorded in Table 32. The only hyperparameter value that remains consistent across all three models is minimum_samples_leaf=1. Both tuned models selected maximum features to be none, meaning there is no limit on the size of the feature subsample in each estimator. Therefore, the algorithm is free to include any number of genes in each split of a tree in the random forest.

Neither the Grid Search tuned, nor the Bayes Search tuned models result in an improvement in the AUC achieved. The Grid Search tuned model has a reduction in sensitivity to identifying the Crohn's Disease class, in comparison to the untuned model (0.61 untuned recall versus 0.50 grid search recall). However, the grid search model is much more sensitive to the ulcerative colitis class (0.63 versus 0.74). Interestingly, this coincides with an almost tenfold increase in the feature importance of *NOD2*, when comparing the untuned model to the Grid Search tuned model (Figure 47A and Figure 47B). This trend in the sensitivity at which the model can identify each class is similar in the Bayes Model, but to a lesser extent. This coincides with a fivefold increase in the importance of *NOD2* (Figure 47A and Figure 47C). Of the top 10 most important genes, three in the untuned model do not appear in the tuned models: *TNS1*, *TNC*, *HTT*. The top 10 important genes stay the same for the tuned models (although the order changes), except for *WDFY4* in the Grid Search tuned model, which is replaced by *P2RX7* in the Bayes Search tuned model.

Table 32 Disease subtype classifier results on the test data for the untuned ML model, the tuned with Grid Search ML model ,and the tuned with Bayes Search ML model (features = 739).

UNTUNED					**TUNED WITH GRID SEARCH**					**TUNED WITH BAYES SEARCH**				
Hyperparameters Selected					**Hyperparameters Selected**					**Hyperparameters Selected**				
Max Depth		None			Max Depth		5			Max Depth		27		
Max Features		sqrt(n_features)			Max Features		None			Max Features		None		
N Estimators		10,000			N Estimators		1000			N Estimators		250		
Min Samples Leaf		1			Min Samples Leaf		1			Min Samples Leaf		1		
Min Samples Split		2			Min Samples Split		5			Min Samples Split		4		
Machine Learning Results					**Machine Learning Results**					**Machine Learning Results**				
	Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1
CD	0.90	0.61	0.63	0.73	CD	0.92	0.50	0.74	0.64	CD	0.92	0.57	0.71	0.70
UC	0.22	0.63	0.61	0.33	UC	0.20	0.74	0.50	0.32	UC	0.22	0.71	0.57	0.34
Average	0.80	0.61	0.63	0.67	Average	0.81	0.53	0.71	0.60	Average	0.81	0.59	0.69	0.65
AUC	0.67				AUC	0.652				AUC	0.656			
Top 10 Genes	*NOD2, DNAH12, TNS1, WDFY4, P2RX7, SPATS2L, TNC, HTT, E2F4, NFATC1*				Top 10 Genes	*NOD2, GZMA, NFATC1, E2F4, HHAT, GALC, ATM, SPATS2L, DNAH12, WDFY4*				Top 10 Genes	*NOD2, E2F4, SPATS2L, DNAH12, NFATC1, GALC, GZMA, ATM, HHAT, P2RX7*			

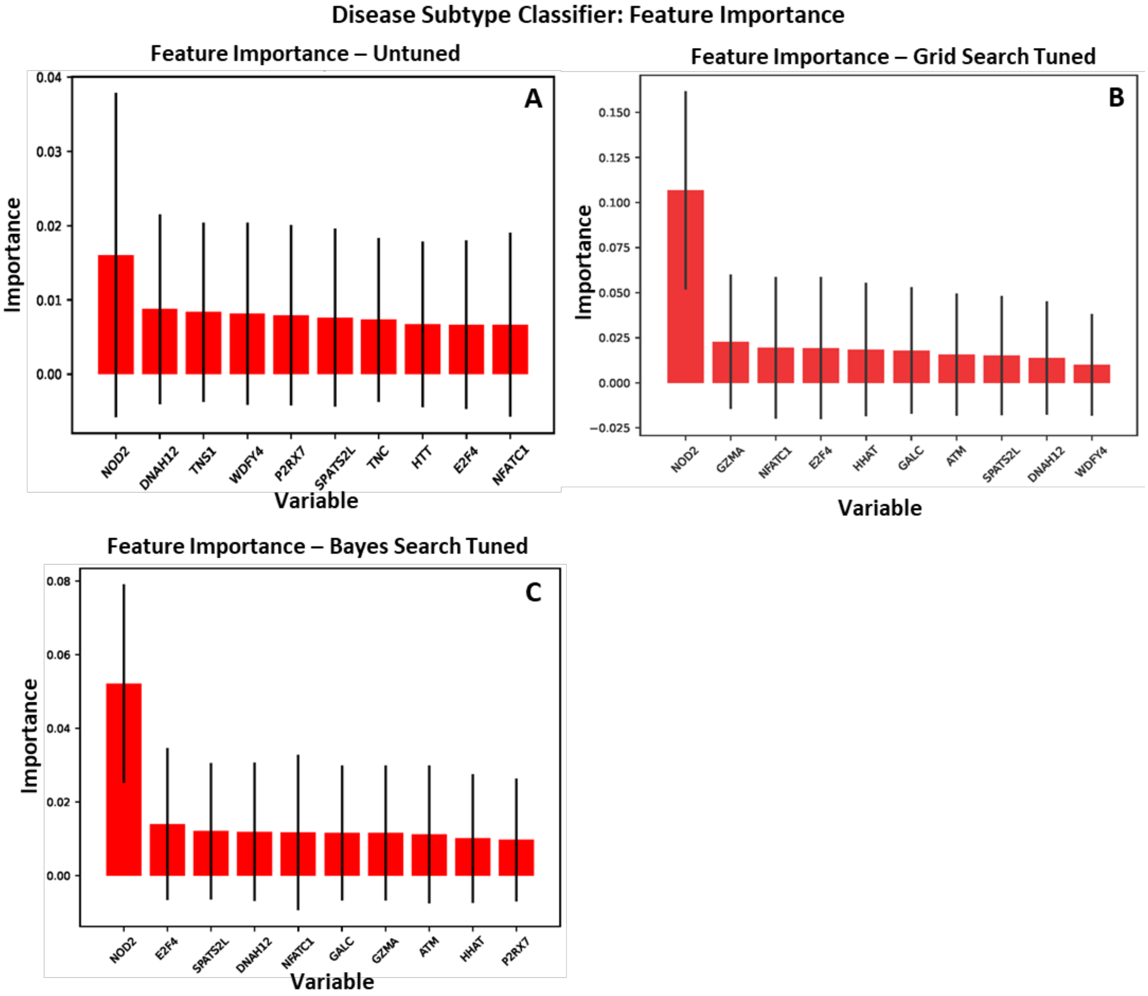


Figure 47 The relative feature importances of the trained random forest using the autoimmune gene panel for the A) untuned model, B) model with hyperparameters tuned by Grid Search, C) model with hyperparameters tuned by Bayes Search

A comparison of the SHAP values of the untuned subtype classifier and the Bayes Search tuned classifier, both utilising the autoimmune gene panel, is shown in Figure 48. Genes determined to have SHAP values that impacted subtype discrimination remained the same, with the exception of *ABCA1*, *IL31RA* and *TRIM63*. In the model tuned using Bayes Search, there are considerably higher SHAP values for *NOD2*, both for the long tail of higher GenePy scores with positive SHAP values contributing to CD classification, and the negative SHAP value cluster which contributes to UC classification.

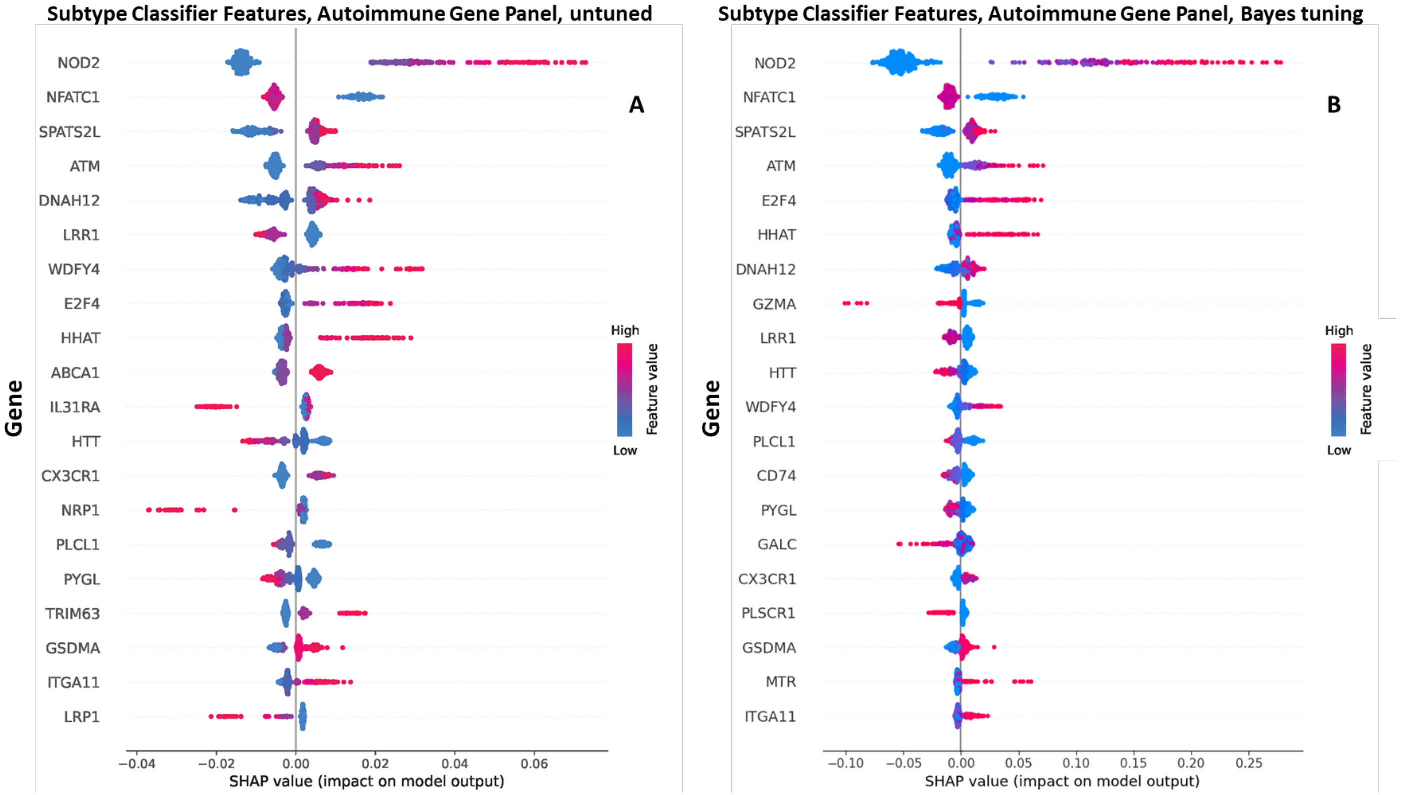


Figure 48 SHAP values for top discriminatory genes, for the disease subtype classifier utilising the autoimmune gene panel. A) Untuned model as shown in Chapter 5; B) after hyperparameter tuning performed using Bayes Search. A high feature value is equivalent to a high CADD score and vice versa. A positive SHAP value indicates the feature makes a contribution to the positive class, which was coding as presence of a stricture.

6.3.2 Stricturing endotype classifier

6.3.2.1 Optimal gene set and patient data to include

The literature search identified 1,023 genes for inclusion in the stricturing gene panel. Some studies listed specific genes that contained implicated variants, such as *NOX4* being identified as protective of fibrotic disease [473], and mouse models showing mutations in IL33 and ST2 promoted intestinal fibrosis [474]. Other studies suggested full pathways as implicated in the development of a stricturing endotype; these included the JAK-STAT, NOD and Nfr2-ARE signalling pathways. The full list of identified genes and their source(s) are detailed in the Supplementary Files. This supplementary information also documents where genes were implicated in stricturing by multiple sources.

A clinical follow-up threshold for the not-stricturing group of Crohn’s disease patients was required for one of the patient datasets. There is an absence of clear clinical guidance on how many years of follow-up would be required in order to determine that a patient would not develop the stricturing endotype. This threshold was arbitrarily set to maximise the follow-up time in the not-stricturing group, while not reducing the sample size. This dataset is imbalanced, with stricturing being the minority class. Due to this, an 8-year follow-up threshold could be set for the not-stricturing group while retaining the same sample size for the balanced training data as in Chapter 5 (136 stricturing, 136 not-stricturing). Roughly equal numbers of stricturing and not-stricturing patients were included in the testing data. The distribution of clinical follow-up time for the stricturing and not-stricturing patient groups is visualised in Figure 49, annotated with the 8-year follow-up threshold.

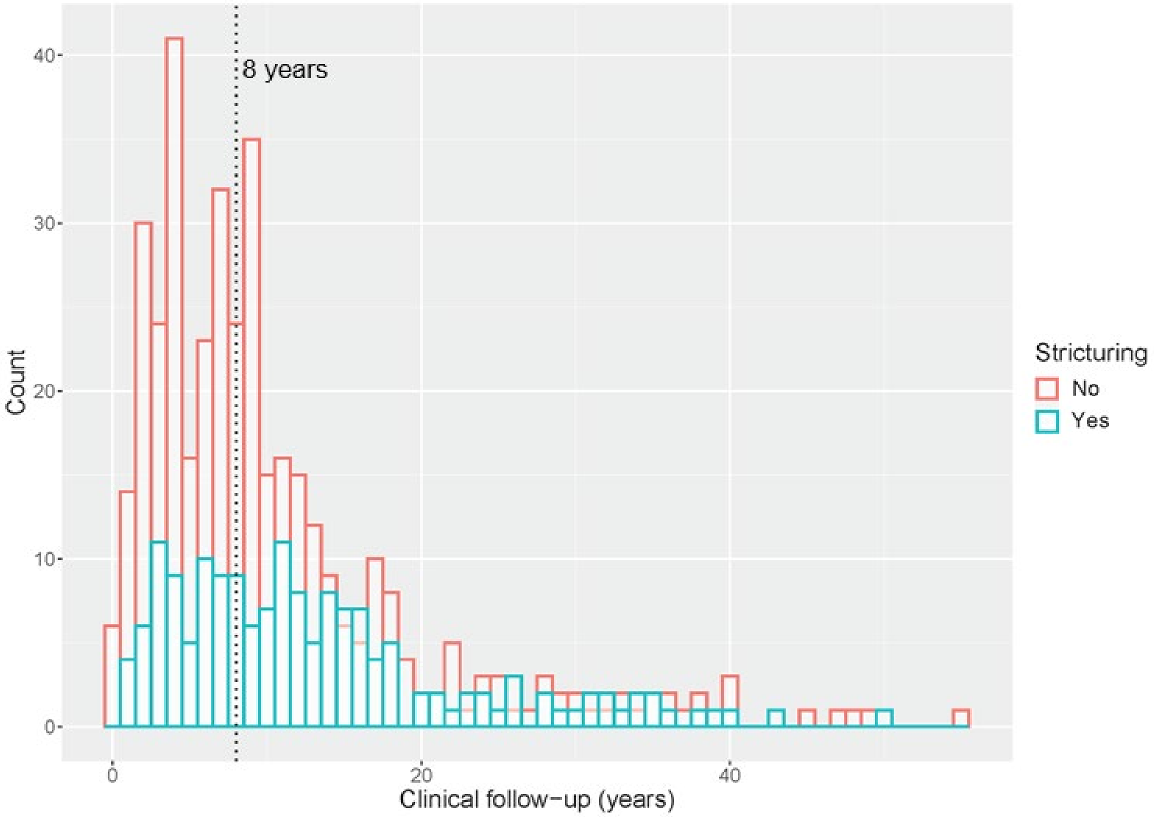


Figure 49 Histogram of years of clinical follow-up for the stricturing and not-stricturing patient groups (n=553, as some CD patients did not have follow-up time data available).

An important aspect to take into account when considering sub-setting the patient data that goes into the stricturing endotype model, is how this might affect the distribution of the data with regards to age of diagnosis. As paediatric patients have, in general, had less time to develop stricturing disease behaviour, the data could potentially be skewed towards the patients with an adult age of onset, who have had more time to develop the endotype. Further, these patients may have been diagnosed in a time prior to the change in treatment approach towards earlier use of biologic therapies, which have been suggested to delay disease progression to stricturing in paediatric patients [475]. In Table 33, the breakdown of the number of patients with a paediatric age of onset, in each class, for the two different datasets, is shown. The follow-up time filter for the not-stricturing class means the percentage of patients with paediatric onset in the stricturing and not-stricturing classes is even. Further clinical detail for the dataset with all data, and the dataset filtered by follow-up time are shown in Table 34.

Table 33 Numbers of patients in each class and the percentage of paediatric onset CD (< 18 yo) per classifier category, depending on the patient set used.

Dataset	**Stricturing Class (% paediatric onset)**	**Not-stricturing Class (% paediatric onset)**	**Total Data (% paediatric onset)**
All Data	170 (42%)	419 (61%)	589 (55%)
Filtered Data (not-stricturing class > 8 yrs clinical follow-up)	170 (42%)	193 (44%)	363 (43%)

Table 34 Clinical characteristics of the full dataset (as used in Chapter 5), and the filtered data, which imposes an 8 year follow-up requirement on individuals in the not-stricturing endotype category. Age at diagnosis information was unavailable for two patients in the full data, and one patient in the filtered data.

		All Data		**Filtered Data (not-stricturing class > 8 yrs clinical follow-up)**	
		Paediatric IBD (< 18 yrs)	**Adult IBD (≥18 yrs)**	**Paediatric IBD (< 18 yrs)**	**Adult IBD (≥18 yrs)**
N		332	255	157	205
Median age at diagnosis (range)		13 (1-17)	31 (18-82)	13 (1-17)	30 (18-82)
Stricturing Endotype	Yes	71	98	71	98
	No	261	157	86	107
Sex	Male	206	113	90	87
	Female	126	142	67	118

Six different gene panels were used in modelling for the two different patient datasets: I) all genes, II) the autoimmune gene panel, III) the IBD gene panel, IV) the extended NOD-signalling pathway panel, V) the stricturing (inclusive) panel, which also includes panels (III) and (IV), and VI) the stricturing (exclusive) panel, which only includes genes identified in the literature search described above. The gene panels are available in the Supplementary Files. The overlap between gene panels are shown in Figure 50. Not all genes listed in each panel could be included in the modelling, as the generation of GenePy scores is based on a reference database which is not exhaustive. In addition, the genetic data pre-processing steps, in particular removal of genes with invariant GenePy scores, mean that the genes included are also dependent on the patient data included. In Table 35, the number of genes in each panel, for both patient datasets documented in Table 34, is recorded.

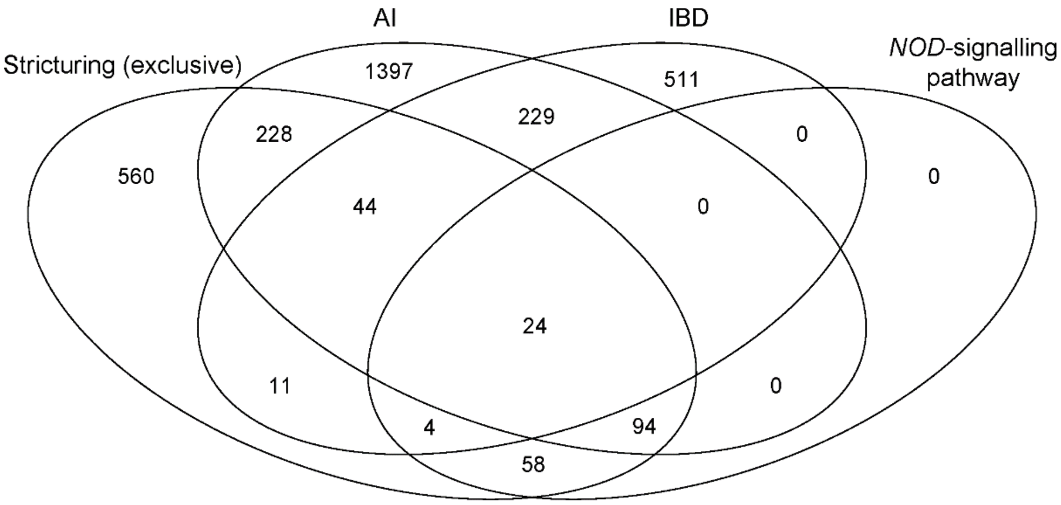


Figure 50 Venn diagram displaying the overlap between the stricturing exclusive panel, the autoimmune gene panel, the IBD gene panel, and the extended NOD-signalling pathway (as labelled from left to right on the diagram. The stricturing (inclusive) panel is not included for clear visualisation, as this includes the genes from the first three panels listed. As the extended NOD-signalling pathway panel contains no genes unique to this panel, the stricturing (inclusive) panel also contains within it the NOD-signalling pathway gene panel.

Table 35 Number of Genes in each of the six gene panels used for random forest modelling of the stricturing endotype. Records total panel genes, genes for which GenePy scores were available, genes after pre-processing for all patient data (n=589), and genes after pre-processing for the patient data where not-stricturing patients are only included if they have over 8 years of follow-up (n=363).

Gene Panel	**Total Genes**	**Genes with GenePy Scores**	**Genes after pre-processing (all patient data)**	**N after pre-processing (patient data with follow-up cut-off)**
All genes	15,669	15,669	14,342	13,490
Autoimmune gene panel	2,017	1,598	1,484	1,397
IBD gene panel	821	499	467	439
Extended NOD signalling pathway panel	180	144	132	121
Stricturing (inclusive) panel	3,155	2,368	2,207	2,085
Stricturing (exclusive) panel	1,023	847	795	754

After patient data pre-processing, 589 CD patients were included in modelling, with 272 patients (136 stricturing, 136 not-stricturing) in the training dataset, and 317 patients (34 stricturing, 283 not-stricturing) in the testing dataset. In Table 36, the test set results of the random forest classifier for the six different gene panels are detailed. For these experiments, the classifier with the highest performance was the model using all available genes (AUC 0.63), and the second-best performing classifier used the NOD signalling pathway genes (AUC 0.58). The classifier that utilised all genes was more accurate in positively identifying patients in the stricturing class (sensitivity 0.59), in comparison to the not-stricturing class (sensitivity 0.54). This was the opposite for the classifier using the NOD signalling pathway panel, which had the highest sensitivity for the not-stricturing class (0.60). There was little overlap in the top 10 genes of each model, which was particularly surprising when considering the autoimmune panel, IBD panel, and stricturing panel (inclusive), as the former two gene panels are included in the latter. The most overlap in the top 10 genes occurred between the two stricturing panels, with three of the overlapping genes belonging to the collagen family. In total, eight genes appear in the top 10 genes of two classifiers, including *NOD2*, *DOCK8* and *CNTRL*. Only two genes appeared in the top 10 genes of three classifiers: *P2RX7* and *GC*.

Table 36 Random forest results for the classification of CD patients by stricturing endotype using different gene panels. All metrics from algorithm performance on the testing dataset (NFS = number of features selected, NS = not-stricturing class, S = stricturing class)

All genes					Autoimmune gene panel					IBD gene panel				
No. Features	534				No. Features	411				No. Features	284			
	Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1
NS	0.92	0.54	0.59	0.68	NS	0.9	0.51	0.53	0.65	NS	0.89	0.45	0.56	0.60
S	0.13	0.59	0.54	0.22	S	0.12	0.53	0.51	0.19	S	0.11	0.56	0.45	0.18
Average	0.83	0.54	0.58	0.63	Average	0.82	0.51	0.53	0.60	Average	0.81	0.46	0.55	0.55
AUC	0.627				AUC	0.518				AUC	0.551			
Top 10 Genes	*PREX1, CNTRL, MAPT, FAT4, GC, AKR7L, PLCE1, PKD1L3, ACACB, PTPRQ*				Top 10 Genes	*TNS1, P2RX7, SPATS2L, LOXL2, BAZ2B, DNAH12, ANK3, FLT4, SORBS1, WDFY4*				Top 10 Genes	*GC, CNTRL, DOCK8, UTP20, NPC1, GALC, GSDMA, NOD2, ERAP1, CD6*			
NOD-signalling pathway gene panel					**Stricturing gene panel (inclusive)**					**Stricturing gene panel (exclusive)**				
No. Features	103				No. Features	462				No. Features	349			
	Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1
NS	0.92	0.60	0.56	0.73	NS	0.90	0.53	0.53	0.67	NS	0.87	0.52	0.32	0.65
S	0.15	0.56	0.60	0.23	S	0.12	0.53	0.53	0.19	S	0.08	0.32	0.52	0.12
Average	0.84	0.60	0.56	0.68	Average	0.82	0.53	0.53	0.61	Average	0.78	0.50	0.34	0.60
AUC	0.578				AUC	0.536				AUC	0.396			
Top 10 Genes	*P2RX7, NLRP3, NOD2, TP53BP1, PLCB3, NOD1, GPRC6A, RNASEL, IRAK2, MAPK12*				Top 10 Genes	*CNTRL, FAT4, GC, TNS1, COL6A2, DOCK8, COL4A4, BMP1, P2RX7, COL27A1*				Top 10 Genes	*FAT4, COL6A2, COL4A4, COL27A1, P2RX7, BMP1, COL15A1, LAMC3, TNC, DNAH17*			

In Table 37, the ML results of these same gene panels are detailed, but using the patient data with a follow-up threshold of 8 years in the not-stricturing group. The training dataset included 272 CD patients (136 stricturing, 136 not-stricturing), and the testing dataset included 91 CD patients (34 stricturing, 57 not-stricturing). Here, the best performing classifier used the IBD gene panel (AUC 0.63), while the next best performing classifier used the stricturing panel (inclusive), which does contain the IBD gene panel (AUC 0.60). Classifiers using gene panels specifically aimed at identifying the stricturing endotype (both the inclusive and exclusive panels) had the highest sensitivity for identifying not-stricturing patients, although these classifiers had a comparatively poorer performance overall. There was more overlap in the top 10 genes across classifiers in this analysis. The all-genes classifier was the only classifier in this analysis section where no gene from the top 10 was featured in another classifier. Apart from the all-genes classifier, *NOD2* appeared in every other classifier as an important gene. Another gene common to four classifiers was the NOD-signalling pathway gene *P2RX7*. There were 11 genes in total that appear in at least two classifiers, and every gene in the top 10 of the stricturing panel (inclusive) appeared in another classifier. There is slightly more consistency in the top genes selected here, in comparison to the analysis that used all patient data. Overall, 26 genes appear in both Table 36 and Table 37.

When choosing the patient data and gene panel which gave the best model, the overall AUC, as well as the sensitivity with which each class could be identified was considered. Some models had better sensitivity for detecting the stricturing class: all genes using all patient data had a sensitivity of 0.59, as did the autoimmune panel with restricted follow up patient data. Different models were better able to detect the not-stricturing class: the stricturing panel (exclusive) and stricturing panel (inclusive) both with the restricted follow-up patient data identifying the not-stricturing class with a sensitivity of 0.67 and 0.65, respectively. The classifier using all genes and all patient data, and the classifier using the IBD panel with restricted follow-up patient data had the highest AUC scores, and these scores were very similar, at 0.627 and 0.630, respectively. As the IBD panel classifier was better at identifying both the stricturing and not-stricturing class, this panel with the restricted follow-up data was taken forward to hyperparameter optimisation.

Table 37 Random forest results for the classification of CD patients by stricturing endotype, **with patient follow up in not-stricturing group > 8 years**, using different gene panels. All metrics from algorithm performance on the testing dataset (NS = not-stricturing class, S = stricturing class).

All genes					**Autoimmune gene panel**					**IBD gene panel**				
No. Features	554				No. Features	411				No. Features	277			
	Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1
NS	0.63	0.47	0.53	0.54	NS	0.70	0.56	0.59	0.62	NS	0.69	0.58	0.56	0.63
S	0.38	0.53	0.47	0.44	S	0.44	0.59	0.56	0.51	S	0.44	0.56	0.58	0.49
Average	0.53	0.49	0.51	0.50	Average	0.60	0.57	0.58	0.58	Average	0.60	0.57	0.57	0.58
AUC	0.554				AUC	0.560				AUC	0.630			
Top 10 Genes	*DPYD, PREX1, CSMD3, PTPRQ, ALG12, PKD1L3, PATJ, THOC6, SERPINB11, USP53*				Top 10 Genes	*TNS1, ACIN1, BAZ2B, NOD2, P2RX7, IFIH1, MAP4, TNNI2, PCSK5, SPATS2L*				Top 10 Genes	*GC, DOCK8, NOD2, TTC7A, IFIH1, NPC1, ERAP1, CD6, CNTRL, UTP20*			
NOD-signalling pathway gene panel					**Stricturing gene panel (inclusive)**					**Stricturing gene panel (exclusive)**				
No. Features	105				No. Features	460				No. Features	358			
	Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1
NS	0.62	0.54	0.44	0.58	NS	0.70	0.65	0.53	0.67	NS	0.66	0.67	0.41	0.66
S	0.37	0.44	0.54	0.40	S	0.47	0.53	0.65	0.50	S	0.42	0.41	0.67	0.42
Average	0.53	0.51	0.48	0.51	Average	0.61	0.60	0.57	0.61	Average	0.57	0.57	0.51	0.57
AUC	0.538				AUC	0.596				AUC	0.547			
Top 10 Genes	*P2RX7, NLRP3, NOD2, NOD1, GPRC6A, TP53BP1, PLCB3, RNASEL, ERBIN, GBP3*				Top 10 Genes	*TNS1, LAMC3, ACIN1, BAZ2B, BMP1, P2RX7, COL6A2, DOCK8, NOD2, COL23A1*				Top 10 Genes	*LAMC3, BMP1, FAT4, P2RX7, NOD2, COL6A2, DNAH17, COL6A6, COL23A1, FKBP10*			

6.3.2.2 Hyperparameter tuning

As for the disease subtype classifier tuning, Table 38 details the values that were trialled for each hyperparameter in each tuning method (**individual hyperparameter tuning**, **Grid Search**, **Bayes Search**). Different values were selected for the Grid Search tuning, informed by the results of individual hyperparameter tuning.

Table 38 Hyperparameter default values, and values tested during individual, grid search and Bayes hyperparameter tuning for the stricturing endotype classifier

Hyperparameter Name	**Default**	**Values Tested – Individual Hyperparameter Tuning**	**Values Tested – Grid Search**	**Values Tested - Bayes**
n_estimators	100	100, 250, 500, 750, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000	2000	100, 250, 500, 750, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000
max_depth	None	1-30, None	5, 7, 11, 12, 15, None	1 - 30, None
min_samples_split	2	2, 3, 4, 5, 6,7 8, 9, 10	2,4,5,6	2, 3, 4, 5, 6
min_samples_leaf	1	1, 2, 3, 4, 5, 6, 7, 8, 9, 10	1,2,3,4	1, 2, 3, 4, 5, 6
max_features	sqrt	sqrt(n_features), log2(n_features), None	sqrt(n_features), log2(n_features), None	sqrt(n_features), log2(n_features), None

As in Section 6.3.1.1, each hyperparameter was individually tuned (Figure 51). Similar to results for the disease subtype classifier, differences between the minimum and maximum values attained by hyperparameter values were within 0.1 of each other. Maximum cross-validated balanced accuracies rarely surpassed 0.55 when changing hyperparameter values, with the best single hyperparameter change being max_depth=12 (Figure 51E). The number of estimators does exhibit the expected trend of increase followed by plateau, however there is a dip when the number of estimators is between 1,000 and 4,000 (Figure 51B). The cross-validated balanced accuracy score for the number of estimators also has a particularly narrow range in comparison to the other hyperparameters. The balanced accuracy increases as the minimum number of samples per split increases (Figure 51A). A similar trend is not seen for the minimum samples per leaf hyperparameter (Figure 51D).

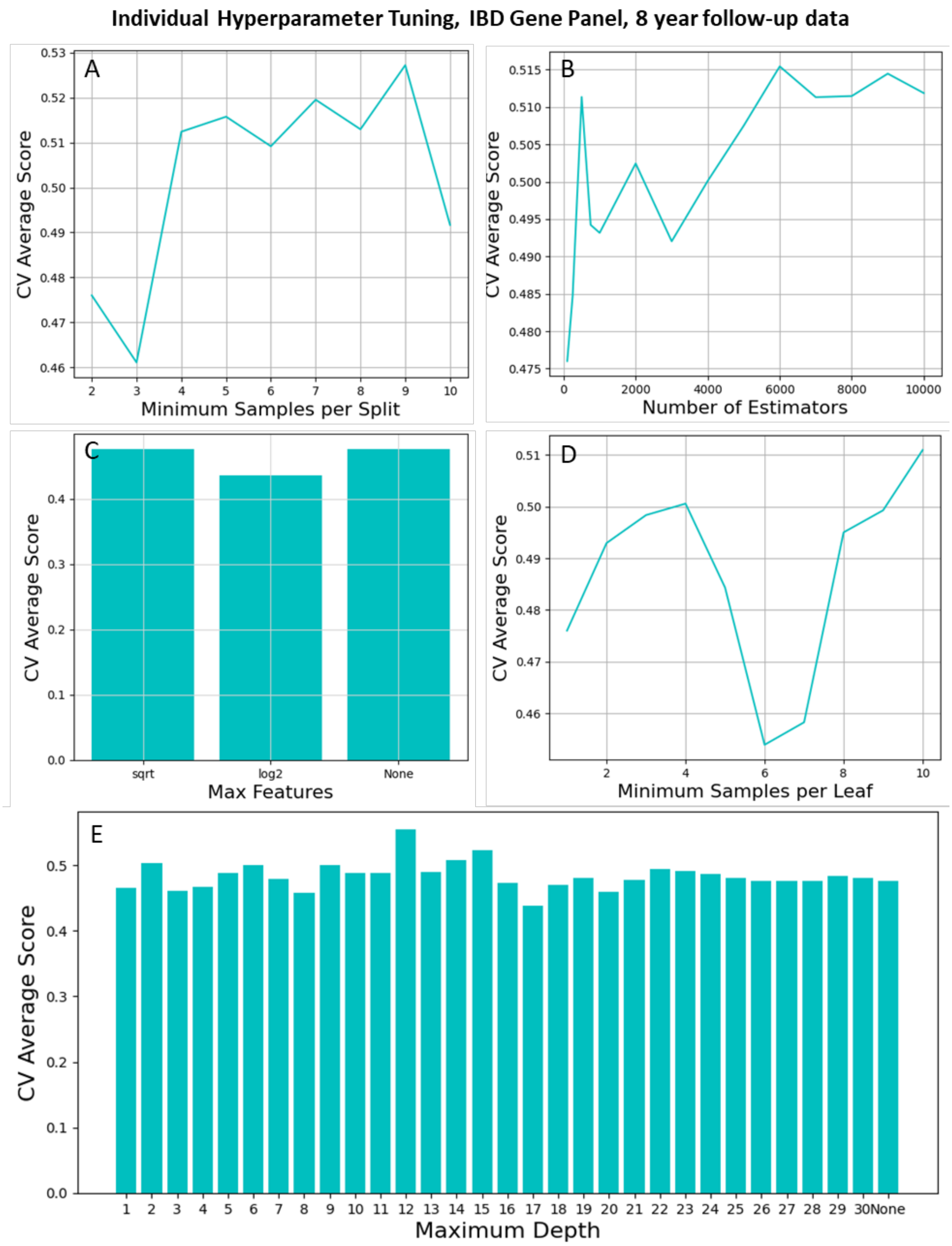


Figure 51 Average balanced accuracy score across the 7 cross-validation (CV) folds for individually tuned hyperparameters (stricturing endotype classifier).

The grid search to tune all hyperparameters simultaneously was performed in a nested cross-validation scheme, using 5-fold inner cross-validation and 7-fold outer cross-validation. Originally, the number of estimators was also going to be optimised. However, with the chosen values of 2000, 6000, 8000, and 10,000 estimators, this became too time intensive, as it comparatively takes much longer to train a random forest with 10,000 trees than one with 100. The time taken to complete the model fitting with the first combinations of hyperparameters was monitored and extrapolating this time out for all hyperparameter combinations indicated that completion of the grid search would take upwards of 84 hours. Therefore, the number of estimators was set to 2000. The chosen hyperparameters for each inner cross-validation fold and the corresponding average balanced accuracy across five folds, and the balanced accuracy in each of the outer cross-validation folds can be viewed in Table 39. The average balanced accuracy across all 7 outer folds was 0.450 (standard deviation 0.060). There was minimal variation in the hyperparameters chosen, with only two values chosen for every hyperparameter apart from minimum samples per leaf (3 chosen). The optimal hyperparameters were chosen according to the balanced accuracy of the corresponding outer cross-validation fold. Here, folds 1, 3 and 7 achieved a joint highest balanced accuracy of 0.500 in the outer fold. Therefore, hyperparameters were chosen based on the average inner cross-validation balanced accuracy, although some hyperparameters were the same across these three folds. The hyperparameters chosen were: maximum depth = 5, maximum features = sqrt, minimum samples per leaf = 1, minimum samples per split = 2, and number of estimators = 2000.

Table 39 Hyperparameter tuning with nested Grid Search for the stricturing endotype classifier. The optimal hyperparameters and corresponding balanced accuracy in the inner cross-validation is given, and the corresponding outer fold balanced accuracy for those hyperparameters to indicate how generalisable these hyperparameters are.

Fold	**Outer Fold Balanced Accuracy**	**Inner CV Balanced Accuracy**	**Optimal Hyperparameters**			
			Max Depth	**Max Features**	**Minimum Samples per Leaf**	**Minimum Samples per Split**
1	0.500	0.517	5	Sqrt	1	5
2	0.364	0.572	7	None	1	2
3	0.500	0.491	5	sqrt	4	2
4	0.357	0.507	5	None	1	2
5	0.488	0.511	5	None	3	2
6	0.444	0.578	5	None	1	2
7	0.500	0.526	5	Sqrt	1	2

The Bayes Search method with the number of iterations set to 60 was then used to optimise the hyperparameters. The same nested cross-validation scheme of 5-folds in the inner cross-validation, and 7-folds in the outer cross-validation was used. The results of this optimisation, with the best inner cross-validation balanced accuracy and corresponding hyperparameters, and the balanced accuracy when applying these hyperparameters to the outer fold test set, can be viewed in Table 40. The only hyperparameter where the value changed for every fold was the number of estimators. The average balanced accuracy across the outer folds was 0.465 (standard deviation 0.050). The best balanced accuracy in the outer fold was 0.521 in fold 6. Therefore, the chosen hyperparameters from Bayes optimisation were: maximum depth = 2, maximum number of features = log2, minimum samples per leaf = 1, minimum samples per split = 4, and 3000 estimators.

Table 40 Bayes Search with nested CV for stricturing endotype classifier. The optimal hyperparameters and corresponding balanced accuracy in the inner cross-validation is given, and the corresponding outer fold balanced accuracy for those hyperparameters to indicate how generalisable these hyperparameters are.

Fold	**Outer Fold Balanced Accuracy**	**Inner CV Balanced Accuracy**	**Optimal Hyperparameters**				
			Max Depth	**Max Features**	**Minimum Samples per Leaf**	**Minimum Samples per Split**	**Number of Estimators**
1	0.50000	0.5100636	2	log2	5	2	1000
2	0.50000	0.5104078	2	log2	1	2	750
3	0.440579	0.5098666	2	None	4	6	5000
4	0.3599439	0.5292156	5	log2	2	2	500
5	0.4722222	0.5238714	13	None	1	4	4000
6	0.5217391	0.5227108	2	log2	1	4	3000
7	0.4637681	0.526412	4	None	1	4	100

6.3.2.3 Application of optimal hyperparameters to random forest modelling

Upon selection of optimal hyperparameter values by Grid Search and Bayes Search methods, the random forest was trained with these two hyperparameter sets using the whole training set. A comparison of untuned and tuned random forest model results on the test set is recorded in Table 41. The only hyperparameter value that remained constant across all three model iterations was min_samples_leaf=1. The model using Grid Search tuned hyperparameter values also had the same values as the untuned model for the maximum features, and minimum samples per split hyperparameters. Aside from the minimum samples per leaf, there were no shared hyperparameter values between the Grid Search and Bayes Search tuned models. Both the Grid Search and Bayes Search tuned models make modest improvements on the untuned AUC of 0.63 (0.65 and 0.66 AUC, respectively). The Grid Search tuned model showed no improvement in sensitivity to the not-stricturing class, but does make a modest improvement on stricturing class sensitivity (improves from 0.56 in untuned, to 0.59). The Bayes Search tuned model shows this same stricturing class sensitivity improvement, but also improves on the sensitivity to the not-stricturing class by a greater margin (0.58 in untuned model, improved to 0.65). In both tuned models, the feature importances of the top 10 genes has decreased (Figure 52). This suggests more genes are being valued as of equal importance in these models, in comparison to the untuned model. The gene importances decrease to a greater extent for the Bayes Search tuned model (Figure 52C). All three models feature the same 4 genes in the top 5 genes by importance: *GC*, *TTC7A*, *NOD2* and *IFIH1*, although their order varies. Additionally, *UTP20* also features in every model. The genes *ERAP1* and *CNTRL* are unique to the untuned model, and the Grid and Bayes Search tuned model have 8 genes in common, with *NPC1* and *CD6* in the Grid Search tuned model being replaced by *MEI1* and *THBS3* (the latter two genes being unique to the Bayes Search tuned model).

Table 41 Stricturing classifier results on the test data for the untuned ML model, the tuned with Grid Search ML model ,and the tuned with Bayes Search ML model (features = 277). Uses the IBD gene panel, and patient data where the not-stricturing group has > 8 years follow-up.

UNTUNED					**TUNED WITH GRID SEARCH**					**TUNED WITH BAYES SEARCH**				
Hyperparameters Selected					**Hyperparameters Selected**					**Hyperparameters Selected**				
Max Depth		None			Max Depth		5			Max Depth		2		
Max Features		sqrt(n_features)			Max Features		sqrt(n_features)			Max Features		log2(n_features)		
N Estimators		10,000			N Estimators		2,000			N Estimators		3,000		
Min Samples Leaf		1			Min Samples Leaf		1			Min Samples Leaf		1		
Min Samples Split		2			Min Samples Split		2			Min Samples Split		4		
Machine Learning Results					**Machine Learning Results**					**Machine Learning Results**				
	Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1		Precision	Recall	Specificity	F1
NS	0.69	0.58	0.56	0.63	NS	0.70	0.58	0.59	0.63	NS	0.73	0.65	0.59	0.69
S	0.44	0.56	0.58	0.49	S	0.45	0.59	0.58	0.51	S	0.50	0.59	0.65	0.54
Average	0.60	0.57	0.57	0.58	Average	0.61	0.58	0.58	0.59	Average	0.64	0.63	0.61	0.63
AUC	0.630				AUC	0.649				AUC	0.660			
Top 10 Genes	*GC, DOCK8, NOD2, TTC7A, IFIH1, NPC1, ERAP1, CD6, CNTRL, UTP20*				Top 10 Genes	*TTC7A, NOD2, GC, IFIH1, DOCK8, UTP20, KPNA7, NPC1, CCNL2, CD6*				Top 10 Genes	*TTC7A, IFIH1, KPNA7, NOD2, GC, UTP20, MEI1, CCNL2, THBS3, DOCK8*			

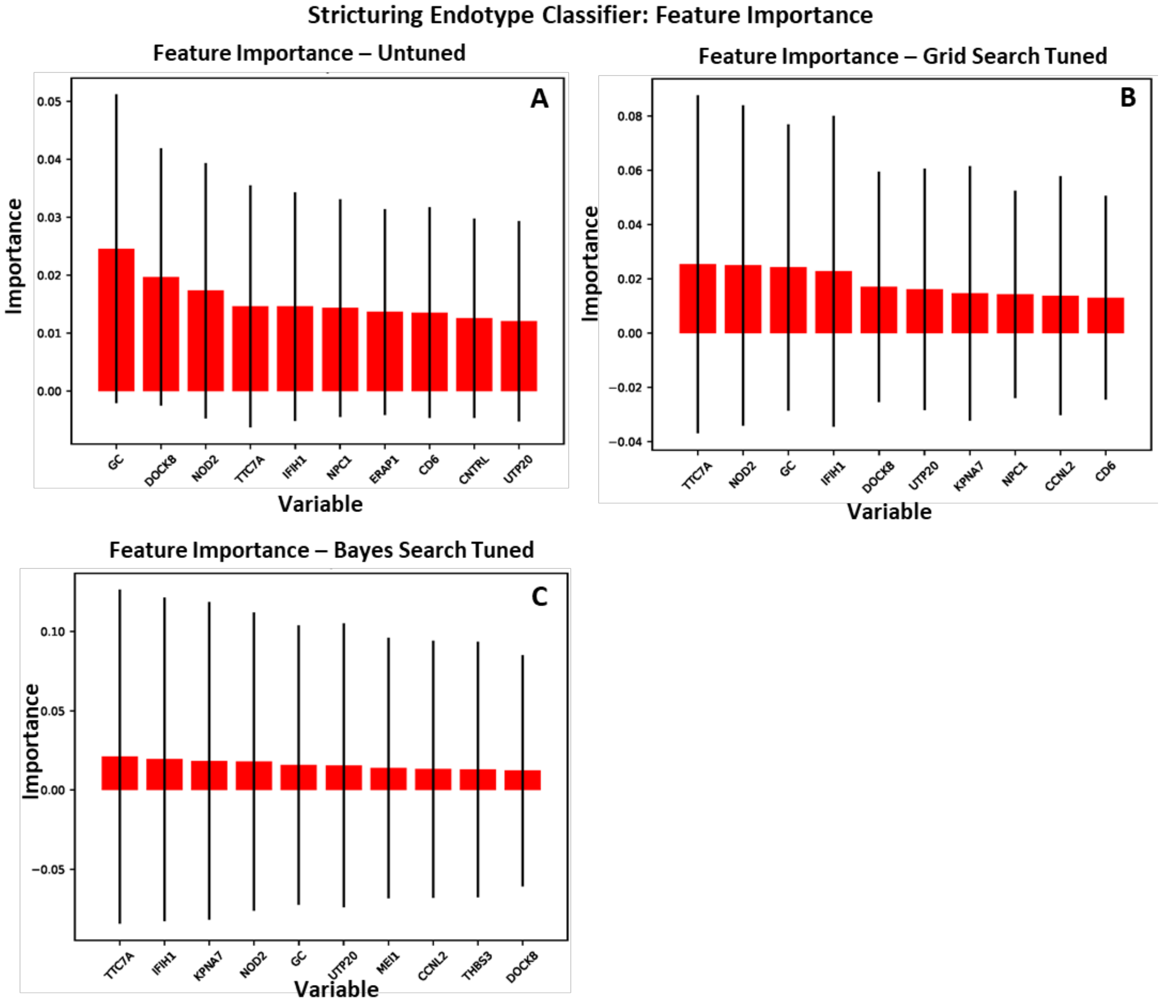


Figure 52 The relative feature importances of the trained random forest using the IBD gene panel with 8-year follow-up on not-stricturing patients for the A) untuned model, B) model with hyperparameters tuned by Grid Search, C) model with hyperparameters tuned by Bayes Search

The SHAP values of the Bayes Search tuned stricturing endotype classifier, utilising the IBD gene panel, are shown in Figure 53. The magnitude of SHAP values observed was small. Variation in some genes appeared to be protective against the development of stricturing endotype (high GenePy scores correlating with a negative SHAP value), for example for *TTC7A*, *UTP20* and *IFIH1*. Higher GenePy scores in other genes, such as *NOD2*, *CCNL2*, and *TGFB1* led to positive SHAP values, indicating variation in these genes contributes to formation of a stricturing endotype. In general it appears that many genes had clusters of low GenePy scores, with only a few datapoints with high GenePy scores. In general, the feature values in Figure 53 appear to result in a less continuous distribution of SHAP values in comparison to the SHAP values achieved in the disease subtype ML modelling.

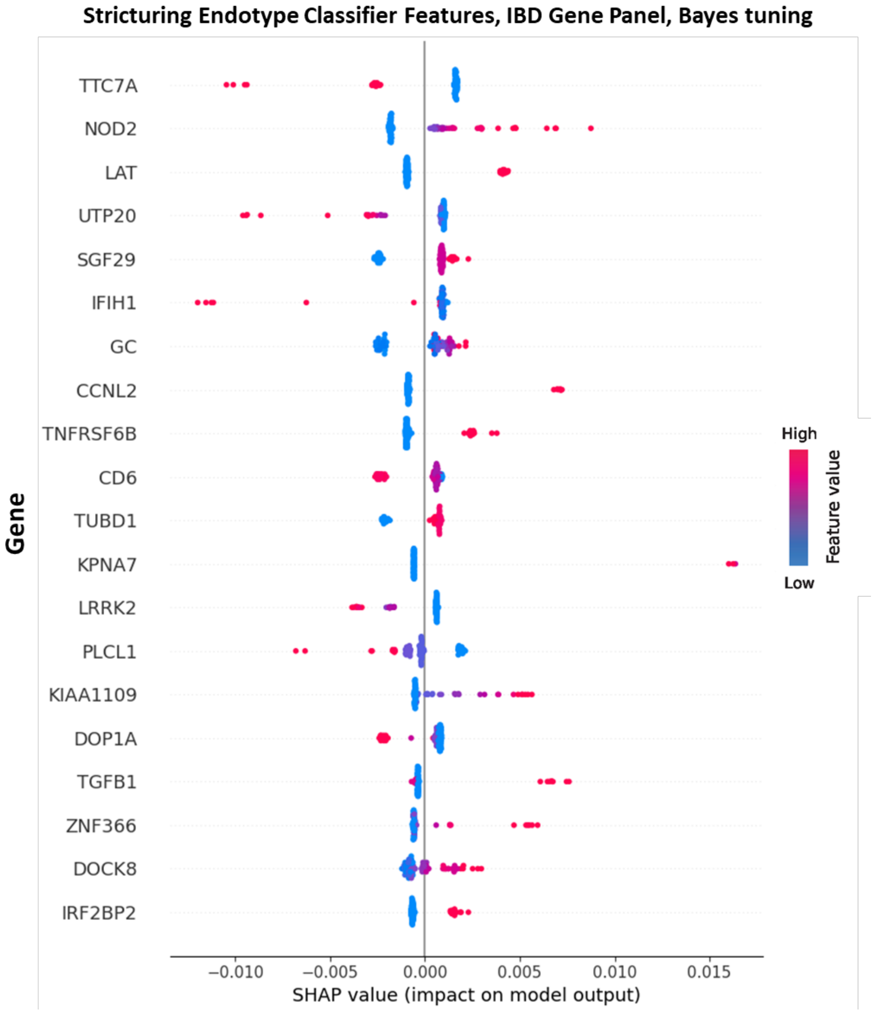


Figure 53 SHAP values for top discriminatory genes, for the stricturing endotype classifier utilising the IBD panel, with hyperparameter tuning performed using Bayes Search. A high feature value is equivalent to a high CADD score and vice versa. A positive SHAP value indicates the feature makes a contribution to the positive class, which was coded as the presence of a stricture.

A pathway analysis was conducted for features chosen by the stricturing endotype classifier when using the IBD gene panel. As in the pathway analysis performed on the chosen features for the subtype classifier which used the autoimmune gene panel, it was known that the starting gene panel for the stricturing endotype classifier was already enriched with pathways related to the immune system and IBD. From 277 genes selected during stricturing endotype classification, 124 pathways were significantly enriched (adjusted p-value < 0.05), and of these 110 were also enriched in the pathway analysis of all the genes included in the IBD gene panel, according to the adjusted p-value. More useful information was determined to be those pathways which were no longer enriched after feature selection, shown in Table 42. Pathways no longer enriched included the TGF-β signalling pathway, and IL-17 signalling pathway.

Table 42 Pathways excluded from the stricturing endotype classifier that utilises the IBD gene panel, and was tuned with Bayes Search. Pathways listed here were enriched in the Enrichr pathway analysis of the IBD panel genes after pre-processing prior to feature selection, but were **not** enriched in the Enrichr pathway analysis of the gene selected during feature selection for the stricturing endotype ML model.

Pathway term	**Overlap**	**P-value**	**Adjusted P-value**	**Odds Ratio**	**Combined Score**
TGF-β signalling pathway	8/94	1.08E-03	3.95E-03	4.20	28.70
Colorectal cancer	7/86	2.83E-03	9.31E-03	4.00	23.44
Maturity onset diabetes of the young	3/26	1.88E-02	4.20E-02	5.85	23.23
Long-term depression	5/60	1.01E-02	2.61E-02	4.09	18.77
Hippo signalling pathway	10/163	3.28E-03	1.06E-02	2.96	16.91
Intestinal immune network for IgA production	4/48	2.08E-02	4.58E-02	4.08	15.79
ErbB signalling pathway	6/85	1.10E-02	2.76E-02	3.42	15.41
GnRH signalling pathway	6/93	1.66E-02	3.87E-02	3.10	12.71
Apelin signalling pathway	8/137	1.08E-02	2.74E-02	2.80	12.65
Estrogen signalling pathway	8/137	1.08E-02	2.74E-02	2.80	12.65
IL-17 signalling pathway	6/94	1.74E-02	3.96E-02	3.07	12.42
Wnt signalling pathway	9/166	1.12E-02	2.79E-02	2.59	11.61
AMPK signalling pathway	7/120	1.67E-02	3.87E-02	2.79	11.41
Progesterone-mediated oocyte maturation	6/100	2.29E-02	4.91E-02	2.87	10.84

6.4 Discussion

This chapter sought to improve upon the initial modelling results of Chapter 5, where a random forest modelling approach was applied to two clinical classification tasks: predicting whether IBD patients were diagnosed with either CD or UC, and predicting which CD patients had developed stricturing disease behaviour. Optimisation of the modelling consisted of tuning the random forest hyperparameters using two different hyperparameter methods: GridSearchCV and BayesSearchCV implemented in Python. Hyperparameter tuning was only performed using the chosen optimal gene set for the disease subtype classifier, and the stricturing endotype classifier. There can be risks with overfitting when hyperparameter tuning, as throughout the tuning process the model is exposed to the whole training dataset. This can result in data leakage from the hyperparameter tuning into the construction of the subtype or stricturing endotype classifier, and the parameters chosen. This leakage can cause overfitting. By utilising a nested cross validation approach, this data leakage does not occur, and overfitting from the hyperparameter tuning process is minimised. Combinations of hyperparameter values selected by each method were implemented and the performance evaluated. In addition, for the stricturing endotype classification task, optimisation was conducted regarding model input data. Several gene panels were compared, and a clinical follow-up cut-off of 8 years was implemented for the not-stricturing patient group, to be more confident that this was the true clinical status of these patients. Overall, optimised classifiers for the disease subtype model did not perform better than the untuned model but did provide useful insight into how optimisation changed the way the random forest model classified the data. Modelling of the stricturing data after gene and patient data optimisation, and hyperparameter tuning, saw an improvement in model AUCs. These improvements made the stricturing endotype classifier on a par with the disease subtype classifier (0.66 AUC and 0.67 AUC, respectively).

6.4.1 Technical considerations

For both subtype and stricturing classifiers, a Bayes Search method delivered a higher model accuracy in the hyperparameter tuning, and a better AUC when the random forest model was trained and tested using the selected hyperparameters. Generally, Bayes search achieved a higher average balanced accuracy across all folds, with a smaller standard deviation, than the grid search method. In addition, when comparing the inner and outer fold balanced accuracies achieved when tuning, the Bayes Search hyperparameters selected in each inner fold generalised better to their respective outer fold in comparison to the Grid Search. It is possible this is due to different data being present in each fold, as the same data split was not used for both Grid Search and Bayes Search. However, Bayes Search’s generalisability to the outer fold is duplicated for the results of both the subtype and stricturing classifier, which does indicate a trend towards the Bayes Search being a better method. To assess this further, the nested cross-validation could be repeated for different sampling of the training set, randomising the data in each fold to give a clearer picture of which hyperparameter tuning method is more robust. However, nested cross-validation is a computational and time intensive process. With Grid Searches taking between 35 and 84 hours to fit all possible models, repeating this process would be impractical. Even though Bayes Search is less time-consuming, at 8 hours to tune hyperparameters, repetitions of this are still time intensive. In future, this could be implemented by transferring the machine learning pipeline to Iridis 5 and running nested cross-validation processes in parallel.

In the machine learning experiments where there was an attempt to try and assess the importance of *NOD2* to the classifier, it was apparent that in the tuned models where *NOD2* took on a larger proportion of importance, that the AUC was impacted to a greater extent in the tuned model when it was removed as a feature. This highlighted the interconnectedness of the feature selection and hyperparameter tuning for model performance. It is possible that a different feature set as input into the hyperparameter optimisation and subsequent random forest model training would result in different model performance. Although, as there were only modest AUC improvements from the hyperparameter optimisation, it is unlikely that a better performance could have been achieved with a different feature set. A possible solution to this would be to perform feature selection and hyperparameter optimisation concurrently. The downside of this pipeline would be an increase in the computational capacity and time required for optimisation.

6.4.2 Disease subtype classifier

The model optimisation by Grid Search of all the hyperparameters, revealed a tendency towards the construction of less complex individual estimators. The maximum depth was set to 5, as was the minimum samples per split in the Grid Search tuning. This led to a relatively shallow tree. This was somewhat offset by there being no limit to the number of features that could be included for each decision (max_features=None). The Bayes Search tuning created a deeper tree, with maximum depth set to 27. The other hyperparameter values of minimum samples per leaf and split, were set to similar values to the Grid Search (1 and 4, respectively). Bayes Search favoured a smaller forest of 250 estimators versus Grid Search’s 1000. Overall, the generalisation between inner cross-validation and outer cross-validation was good for both these methods, aside from two folds in tuning using Grid Search, which could be due to the limited hyperparameter options for this method.

Although AUCs were very similar for both tuned models, the sensitivity with which these models could identify the different subtypes varied. Classifier sensitivity to individuals with UC is slightly better in the untuned model, and amplified by hyperparameter tuning using either method. In the Grid Search tuned classifier this excellent identification of UC patients came at the expense of the

CD patients, where their detection was no better than chance. The loss in identification of CD patients in the Bayes Search tuned model was minor by comparison. This high sensitivity for UC patients came with a large increase in the importance of *NOD2*, which was more pronounced for Grid Search tuned model. When observing the differences between SHAP values produced from the untuned IBD subtype classifier versus the Bayes Search tuned subtype classifier, the increase in SHAP value for *NOD2* is striking (approximately 0.08 in the untuned classifier, approximately 0.27 in the Bayes Search tuned classifier). Although the hyperparameter tuning does not improve the testing AUC, it does cement *NOD2* as a highly discriminant gene for the subtype classifier. A possible theory for why the signal is so strong for *NOD2* is that this is due to *NOD2*'s implication in both monogenic [5] and polygenic [2, 3] CD.

This combination of the high NOD2 SHAP and feature importance, combined with an increased UC sensitivity, led to an interesting and unexpected finding. This combination was unexpected due to NOD2's strong association with CD. This combination of results suggests that during model training the presence of a genetic signal in *NOD2*, indicating a pathogenicity burden, does not mean the model is more likely to classify a patient as having been diagnosed with CD. Rather, that the absence of a genetic signal in *NOD2* results in a UC diagnosis being considered as more likely by the random forest algorithm. This is reflective of standard CD pathology, as not all CD patients will have deleterious variation in the *NOD2* gene.

6.4.3 Stricturing endotype classifier

A literature review was conducted to attempt to assemble a comprehensive list of genes with the potential to be associated with the development of a stricturing endotype. Therefore, there was a low threshold of evidence for a gene's inclusion, such that as many genes that have the potential to cause stricturing as possible were included. Nevertheless, as is the nature with any literature review, it is possible that there are hitherto undiscovered genes that contribute to the development of a stricture in CD patients which will not be included. However, as Figure 50's Venn diagram illustrates, 560 genes that had not previously been included in any other ML model gene panel were present in the stricturing (exclusive) panel, which is a large uplift in genes that could be potentially connected to development of a stricturing endotype.

While there was significant overlap in the 6 gene panels utilised in modelling, there were several reasons for their inclusion. Firstly, while in Chapter 5 the best AUC was achieved through an agnostic approach (including all genes), the subsequent Enrichr [454] pathway analysis resulted in no significantly enriched pathways. Therefore there was a need to investigate different gene panels, and whether a filtered patient dataset could result in a better performing model. The two panels previously used in Chapter 5: the autoimmune gene panel, and the IBD gene panel, were used for consistency, as a new patient dataset where a clinical follow-up cut-off of 8 years was implemented for the not-stricturing patient group was trialled. Therefore inclusion of these panels allowed for a comparison in results achieved during Chapter 5, and results with the filtered patient data. Secondly, the stricturing (exclusive panel) was utilised to assess if model performance could be improved using a gene panel tailored to this specific endotype. The stricturing (inclusive) panel was generated to incorporate three gene panels together: the stricturing genes, the IBD gene panel, and the autoimmune gene panel. This was the largest panel, and as such there were some concerns about the highly-dimensional nature of this dataset (where the number of features exceeds the number of samples). This was another reason to use both the stricturing (inclusive) and stricturing (exclusive) panels. Opposite to this, was the use of the NOD-signalling gene pathway panel, which contained a small number of genes, and as such was not a highly-dimensional dataset. This pathway had the potential to be an influence on the development of a stricturing endotype, given the already known involvement of *NOD2* in its aetiology [5, 388].

When random forest modelling was performed using several gene panels in combination with the patient data where the clinical follow-up threshold was used, some of these results echoed the relationship observed in the previous classifier between *NOD2* and UC. Results from gene panels with a more general basis – the autoimmune panel and the IBD panel – exhibited similar sensitivities for both stricturing and not-stricturing groups. The two stricturing panels did not produce the best model, but they were both more sensitive to the not-stricturing class in comparison to the rest of the panels (especially in the case of the stricturing exclusive panel). As before, this suggests the absence of a pathogenicity signal from these genes is more beneficial for classifying not-stricturing patients, and not the presence of a signal confirming the stricturing endotype. This suggests the presence of genomic heterogeneity in both the disease subtype and stricturing endotype classifications. The IBD disease subtypes are thought to have subgroups within them based on the different molecular mechanisms causing immune dysregulation. These results suggest this may also be the case for the CD stricturing endotype.

The SHAP values for the Bayes Search stricturing endotype classifier, utilising the IBD gene panel, are sparser than those produced by the stricturing classifier utilising all genes in Chapter 5. This is to be expected given the reduced patient dataset utilised in ML modelling. There was an observed stronger delineation between extremes of GenePy scores, and these extremes were associated with either a positive or negative SHAP value. This was not the case for stricturing endotype modelling in Chapter 5, where lower GenePy scores appeared to be associated with both positive and negative SHAP values for some genes. There appears to be small clusters of high feature values with larger SHAP values (either positive or negative). This may indicate small groups of patients driving discrimination of stricturing and not-stricturing endotypes. This is potentially indicative of genetic heterogeneity within these two classes. From the SHAP values, variation in some genes appears to be protective against stricturing (*TTC7A*, *UTP20*, and *IFIH1*), and other genetic variation appears to contribute to development of a stricturing endotype (*NOD2*, *LAT*, and *CCNL2*). Unlike the IBD subtype classifier, where the size of SHAP values for some genes increased with tuning, the SHAP values observed here remained of the same magnitude as the stricturing endotype modelling conducted in Chapter 5. The pathway analysis revealed some immune signalling pathways key in IBD were excluded by feature selection from the stricturing endotype classifier: TGF- β signalling pathway, and IL-17 signalling pathway. TGF- β is an immunosuppressive cytokine, and TGF- β signalling impairment has been shown to cause spontaneous colitis in mouse models [476]. However, literature does not concord with the ML model, as TGF- β1 expression has been recorded as higher in CD patients with strictures [477, 478]. In fact, TGFB1 is present on the SHAP value plot. This contradictory analysis highlights the caution that must be taken when attempting to construct definitive relationships between genes and the ML modelling outcome.

For hyperparameter tuning by both Bayes Search and Grid Search methods, inner cross-validation balanced accuracies were only slightly better than random chance, and some of these generalised very poorly to the outer fold, resulting in model performances worse than random. This indicates that the sample size in each of the cross-validation folds is too small for the models to learn the data patterns. This is potentially due to combining the small sample with genomic complexity. In this case, it may have been more beneficial to concede the assessment of generalisability that a nested cross-validation approach gives, in favour of increasing the sample size in each fold. Both hyperparameter tuning approaches favoured a shallow tree, with a maximum depth of 5 and 2 selected by Grid Search and Bayes Search, respectively. The maximum features hyperparameter value of log2(number of features) also means the estimators created by the Bayes Search tuning are less complex.

This simpler approach to the construction of random forest estimators yielded the best results, with the Bayes Search model improving most upon the untuned model results. This model once again features a higher sensitivity to the not-stricturing class. The generation of simple estimators only leads to a good performance, and it may be the case that without estimators of equivalent complexity to the dataset, there is a limit to the AUC that can be achieved in this modelling. A potential way to improve the modelling here would be to more fully incorporate the longitudinal data present in the dataset. This could be done with survival analysis models, or methods that combine machine learning and survival analysis. Then the full dataset could be utilised, as a clinical follow-up cut-off would not be necessary, and it would be more flexible to the inclusion of new and updated data, as follow-up time increases, or new patients are recruited.

These results suggest that the development of prognostic classifiers for an endotype such as stricturing disease may be more challenging than first thought. Narrowing the modelling scope to a more specific endotype has probably eliminated some of the genetic complexity behind the groups the model is trying to classify patients into, in contrast to the disease subtype model. This reduction in complexity is partially mitigated by a reduction in sample size, due to the performance of modelling on CD patients only, the not-stricturing patient data being limited by a clinical follow-up threshold of 8 years, and the imbalanced nature of the classification problem. It is likely that within this relatively small dataset, there exists many combinations of genetic variation that could lead to a stricturing endotype. A larger dataset may allow a model to make the connection between patient's genomics and their endotype, as there would be more examples for each group of patients with a similar genomic signal.

Both the disease subtype classifiers and stricturing endotype classifiers have highlighted genes that were influential in classification. A limitation here is that these models did not exhibit excellent performance through their testing AUCs. Therefore, these models would not be expected to generalise across all datasets, and be representative of all patient populations. This makes deciphering a strong link between a model outcome and the set of input genes extremely challenging. The small feature importances observed in both types of classifiers (aside from *NOD2*), and the hundreds of genes selected as input for these ML models, combines with the testing AUCs to suggest it would be unwise to draw conclusions from the genes selected during feature selection for either the IBD subtype classifier, or the stricturing endotype classifier. Instead, a potential conclusion is that these classifiers behave in a similar way to IBD: they are polygenic, with each gene (feature) making a small contribution to discriminating one class from the other.

Chapter 7 Random survival forest for stratification of Crohn's disease patients by stricturing endotype

Chapter summary – in this chapter, a new pipeline for the stratification of Crohn's disease patients by stricturing endotype is constructed and improved. Cox Proportional Hazards modelling and Principal Component analysis is used as feature selection. Machine learning modelling is performed using three gene panels. Random survival forest models were chosen for stratification to incorporate clinical follow up times into stricturing endotype modelling. These models were optimised using Bayesian hyperparameter tuning.

Chapter contributions – Whole exome sequencing data was joint-called by Guo Cheng. All subsequent processing, and transformation into GenePy scores were performed by Imogen Stafford. The IBD gene panel was curated by Guo Cheng and James Ashton. Other gene panels were curated by Imogen Stafford. Clinical stricturing status and follow-up were assessed by Imogen Stafford and James Ashton. Modelling was performed by Imogen Stafford.

Supplementary files can be found at https://doi.org/10.5258/SOTON/D2655

7.1 Introduction

In Chapter 5, stratifying patients by stricturing endotype was approached as a straightforward classification task with a random forest. In Chapter 6, the patient data included in the machine learning classification was changed, and patients were only included in the not-stricturing group if they had at least 8 years of clinical follow-up. This was to account for the time it can take patients to develop a stricturing complication. Here, this idea is taken to its natural conclusion, by transforming this machine learning problem from a classification problem, into a regression problem, and integrating survival analysis techniques into the ML prediction.

The simple way of estimating survival (or time to event in general) is to use a Kaplan Meier estimate, which can be plotted as Kaplan Meier curves. This assesses the differences in survival for a categorical variable, for example treatment type, with a log-rank test measuring the statistical significance. The multivariate version of this modelling is Cox Proportional Hazards (CPH) modelling. With this method, the effect of each independent variable on the event occurrence is assessed, by observing how these variables impact the hazard ratio [479]. Here, the variables in the dataset are assessed for their impact on the ratio between the baseline hazard, and a patient's (or group of patients) own hazard score. As with machine learning methods, a CPH model is usually fitted to a dataset, and then tested on a new dataset. To assess the success of the model, a modified form of AUC is used, called Harrell's concordance index, or the C-index. This metric calculates the proportion of all possible pairs of patients in the data where the CPH model prediction is concordant with outcome. For each pair, at least one patient must have experienced the event. A successful prediction for this clinical problem is defined by whether the model predicted patient 1 would stricture before or after patient 2, and if this matches the outcome in the test data [480]. CPH models are assessed in this way due to the censored data present for all survival problems. The AUC over time can also be used to assess the model, as for a specific time point the problem effectively becomes a classification problem again.

The principles of CPH modelling have also been combined with machine learning algorithms. LASSO and ridge regression penalties have been combined with CPH models to effectively provide feature selection within the model [481]. Support vector machines have included survival analysis principles through uncertainty and weighting. The model is still a binary classification problem, but the classification of samples comes with a weight which includes follow-up time, and represents how likely it is that the event occurred for an individual [482]. In addition, neural networks have been developed that incorporate CPH modelling, such as Cox-nnet. This algorithm uses two layers of a neural network, where the first hidden layer transforms the data in order to model its complex patterns (and as such can be interrogated to determine the important features), and the second layer performs a Cox regression [483]. Random survival forests (RSF) are a regression-based variation of standard random forests used in Chapters 5 and 6. The key difference between a random forest regression and an RSF is the metric that the algorithm takes into account when determining each split in a tree. For a regression, a tree split would be chosen in order to minimise the residual sum of squares. In RSF, a split is chosen to maximise the log-rank test, ensuring the maximal difference in survival between the two daughter nodes [484].

A highly relevant paper to the clinical problem discussed here is the 2021 study by Ungaro et al. [485] of possible blood proteomic markers in paediatric IBD patients. They utilise RSF as a feature selection method, to identify possible blood proteomic markers associated with stricturing and penetrating endotypes. They use a RSF as both feature selection, and to produce a final model. Their model, utilising four protein markers, achieved an AUC of 0.68 (5-fold cross validation with 200 bootstrapped repetitions on one dataset). This highlights the potential a RSF algorithm may have for stratifying patients by this endotype.

This chapter aimed to create a machine learning pipeline that could utilise the time-to-event patient data for the stricturing endotype. The genomic data of paediatric and adult CD patients in the form of GenePy scores is utilised as the input data for stratification. RSF was chosen as the machine learning algorithm because it could make full use of the clinical follow-up data to best exploit the comparatively modest sample size. A neural network-based model would require a larger cohort. Additionally, it is a non-linear method, and therefore is equipped to model the non-linear gene-to-gene interactions thought to be present in genomic data. Several feature selection methods were trialled, first with a focus on utilising follow-up data in feature selection, and later focussing on reducing genomic dataset dimensionality and sparsity. Different gene panels were tested as an input for machine learning as in previous chapters. Model performance was always assessed on a test dataset, to observe how generalisable these types of models were, and assess potential applicability in clinical practice.

7.2 Methods

The whole exome sequencing data was aligned, joint-called, annotated, and GenePy scores generated as in Section 3.4. Clinical follow-up data to determine stricturing status, and time to stricture were collected as described in Section 5.2.1. All downstream modelling was performed in Python (v.3.7).

7.2.1 Initial random survival forest pipeline

Pre-processing of patient data and GenePy scores was performed (CADD Phred cut-off of 15 used, see Chapters 5 and 6). Genes on the remapped Fuentes list of false positive genes and those with no variance in their GenePy scores were removed. Modelling was performed using the extended NOD signalling pathway gene panel (KEGG:hsa04621 and REACT:R-HSA-168638 pathways). Unrelated patients with highly confident European ancestry prediction (confidence > 0.9) were included, as determined by Peddy [443]. Additionally, each patient must have a record of the number of years of clinical follow-up. This last filter caused a modest reduction in the overall dataset in comparison to the dataset used in the classification problems of Chapter 5 and 6.

For the feature selection method that was chosen in this pipeline, the input matrix needed to have linearly independent columns, i.e.

$$\mathrm{G} = \begin{bmatrix} \mathrm{g}_{11} & \mathrm{g}_{12} & \cdots & \mathrm{g}_{1n} \\ \mathrm{g}_{21} & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ \mathrm{g}_{m1} & \cdots & \cdots & \mathrm{g}_{mn} \end{bmatrix} = (g_{mn}) \in \mathbb{R}^{mxn}$$

$$\text{and } \boldsymbol{v}_1 = \begin{pmatrix} g_{11} \\ \vdots \\ g_{m1} \end{pmatrix} \; such\ that\ \mathrm{G} = \begin{bmatrix} | & | & | & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \cdots & \mathbf{v}_\mathrm{n} \\ | & | & | & | \end{bmatrix}$$

where *G* represents a GenePy matrix with m rows and n columns, and *v₁,v₂, …, vₙ* are the set of vectors representing each gene's set of GenePy scores. *G* has *n* linearly independent columns if and only if the solution to the equation,

$$b_{1\boldsymbol{v}_1} + b_{2\boldsymbol{v}_2} + \cdots + b_{n\boldsymbol{v}_n} = 0, \quad \mathrm{b_n} \in \mathbb{R}$$

is trivial,

$$b_1 = b_2 = \cdots = b_n = 0$$

where $b_1, b_2, …, b_n$ are a set of constants.

In a GenePy matrix, linearly dependent columns are unlikely to present as one column being a multiple of another, as the makeup of GenePy scores across genes is very different. Rather, this occurs if a gene is invariant, and two or more vectors (GenePy score gene columns) are null. While in pre-processing these invariant genes are removed, they can re-occur after the data has been split into training and testing datasets, especially if the GenePy matrix is sparse.

In order to ensure that any training dataset (any *k*), for any value of *n*, contained only linearly independent columns, the binomial probability mass function was used:

$$P(x) = \binom{k}{x} p^x (1-p)^{k-x}$$

where *k* is the size of the training set, the probability *p* is the probability of success, and *x* is the number of successes. This calculation is carried out for each gene column. Success is defined as the selection of an individual into the training set with a GenePy score greater than zero. The probability of success is therefore the proportion of individuals for that gene who have a GenePy score greater than zero. However, for a linearly independent column it is not required that all individuals selected have a GenePy score greater than zero; only one individual must have a non-zero score. Therefore, the binomial probability equation becomes,

$$1 - P(x = 0)$$

that is, success is 1 minus the probability that every individual selected has a GenePy score of zero (zero successes). This is the equation for one gene, but this selection process needs to generate a dataset where all columns are linearly independent. The equation then becomes:

$$\prod_{i=1}^{n}\big(1 - P(x = 0)\big), \qquad \text{for} \quad \mathbf{v}_n \in G$$

In practice, this equation is used to reduce the dataset. If the probability of the above equation is less than 0.95, then the algorithm will remove the column for which *p* is smallest. Then the equation will be re-evaluated for those *n-1* columns. The process repeats until the equation result is above the 95% confidence threshold. A complication of this process was how the training set was generated. The training set is not just 75% of the full dataset, but 75% of the minority class (stricturing in this case), with the same number of samples selected from the majority class. This means that *p* is different for each class. Now the equation becomes,

$$\prod_{i=1}^{n}\left(1 - \binom{k/2}{0} p_s^0 (1 - p_s)^{(k/2)-0}\right) \times \prod_{i=1}^{n}\left(1 - \binom{k/2}{0} p_{ns}^0 (1 - p_{ns})^{(k/2)-0}\right), \qquad for \quad \boldsymbol{v}_n \in \mathrm{G}$$

where p_s is the proportion of individuals classified as stricturing with a GenePy score greater than 0, and p_{ns} is the same for the not-stricturing class. As before, the equation was evaluated, and a column removed, until the result was a 95% or above confidence that a dataset of linearly independent columns was chosen. In each cycle, the column for which *p* is smallest is removed, where

$$p = p_s + p_{ns}$$

This gene filtering approach has a distinct function, in comparison to more widely used feature selection processes, such as L1 (LASSO) and L2 (ridge) regularisation techniques. These regularisation methods operate in relation to the machine learning loss function, that is the square of the difference between the predicted outcome, and the true outcome. In contrast, this binomial probability filtering method only uses information regarding how many non-zero values are present per column (gene), in each class, and does not try to infer any prediction from this information. This also means that information leakage is unlikely.

After the dataset was reduced into a subset of linearly independent columns (genes), the dataset was split into training and testing data in a 75:25 ratio, where the split calculation was based on the number of individuals in the minority class (patients with stricturing behaviour). Then the training dataset was scaled using MaxAbsScaler, and this scaling applied to the testing dataset. The feature selection process first involved constructing individual CPH models for every gene on the training dataset using the CoxPHSurvivalAnalysis tool from scikit-survival. The genes were then ranked by their CPH C-index score. Next, BayesSearchCV, from scikit-optimise was used to determine how many features were included in downstream modelling. The Bayes Search did not evaluate the different combinations of features, instead determining a "top *F*" number of features to include as per the rankings established by individual gene CPH model C-index. The number of Bayes Search iterations depended on the number of genes that were input into the feature selection. These features were then used in training the random survival forest (RSF), implemented through scikit survival. The model was evaluated on the test dataset, using a time dependent AUC, average AUC over time, and C-index. Feature weights were obtained using PermutationImportance from the ELI5 package, which includes functions for debugging, and the explanation of machine learning predictions. This pipeline is illustrated in Figure 54 (see Supplementary files for machine learning scripts).

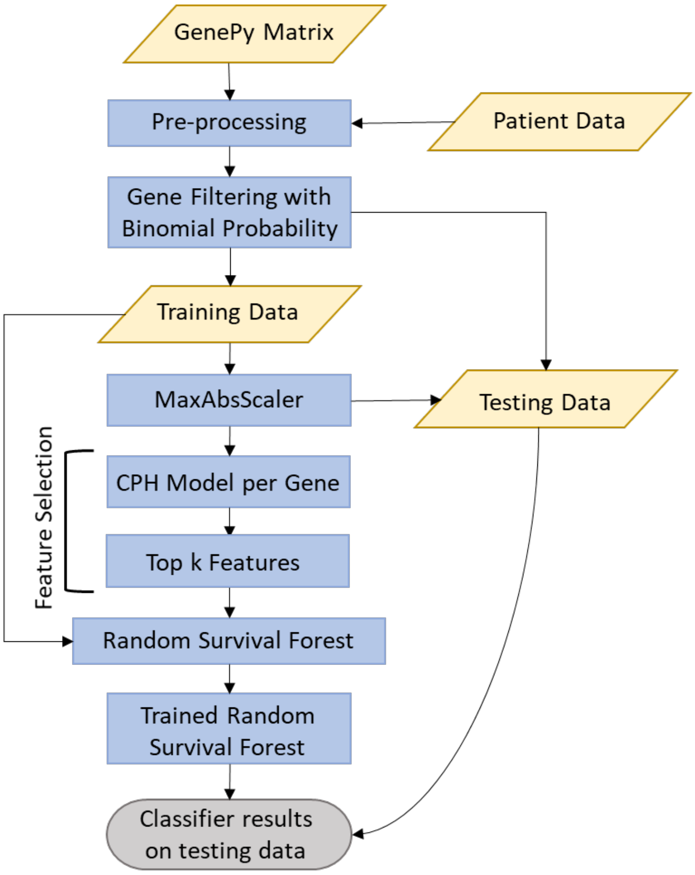


Figure 54 Initial pipeline for random survival forest modelling. Genomic and patient data is pre-processed according to ancestry, relatedness and gene invariance prior to the novel gene filtering method described above, necessary for the feature selection step. Training data is scaled, and this scaling is applied to the test data. Feature selection by ranking CPH model C-index is performed, and the top k features selected by BayesSearchCV. Finally, the RSF algorithm is trained and tested.

7.2.2 Pipeline refinement and final pipeline

Through the implementation of the modelling pipeline in Figure 54, some instabilities were discovered in the modelling, primarily due to the feature selection process. Performing feature selection on a different 90% of the training dataset many times resulted in a different number of "top F" features genes being chosen by BayesSearchCV, due to variations in each gene's CPH C-index. To confirm that this was not due to BayesSearchCV, the process was repeated using GridSearchCV from sci-kit learn [446], which would trial all possible "top F" features.

To stabilise the CPH C-indices that each gene achieved, 90% of the training dataset was subset, and a CPH model for each gene built on this subset. This process was performed 50 times, and the median C-index after different numbers of iterations calculated. The C-indices achieved after different numbers of iterations were assessed, to judge how many were necessary before the C-index for each gene settled to a consistent value (this experiment was performed using the NOD-signalling gene panel). The median C-index was then used to select "top F" features as before. Instead of using BayesSearchCV to accomplish this, a package called kneed was used, which utilises the kneedle algorithm [486]. It determines the point of maximum curvature of a set of points, in this case the median C-index from each gene's CPH models. This process is similar to the use of the elbow method when determining the optimal number of clusters in unsupervised modelling. The knee determined by the algorithm would be the "top F" features for use in the RSF.

In addition, to see if this pipeline could be improved upon, principal component analysis (PCA) was trialled as a potential alternative to CPH modelling. This process involved experiments which used either PCA alone, or in combination with the linearly independent data subset method described above (this was not required to run the PCA, as it was with CPH modelling), and/or the kneed method, for selecting the top principal components (PCs) for inclusion in the RSF.

After feature selection was finalised (for both CPH model, and PCA methods), hyperparameter tuning was performed using the features selected by these methods, in order to produce the optimal model. The hyperparameters tuned in this process were the same as those in the standard random forest: number of estimators, minimum samples per leaf, minimum samples per split, maximum features, and maximum depth (see Section 6.1.1). Tuning was performed using BayesSearchCV in a nested cross-validation scheme (3-fold inner, and 5-fold outer cross-validation) that used the C-index to assess model performance for different sets of hyperparameter values. This, combined with two different feature selection methods, results in two final pipelines that utilise the RSF algorithm to stratify patients by stricturing endotype (Figure 55).

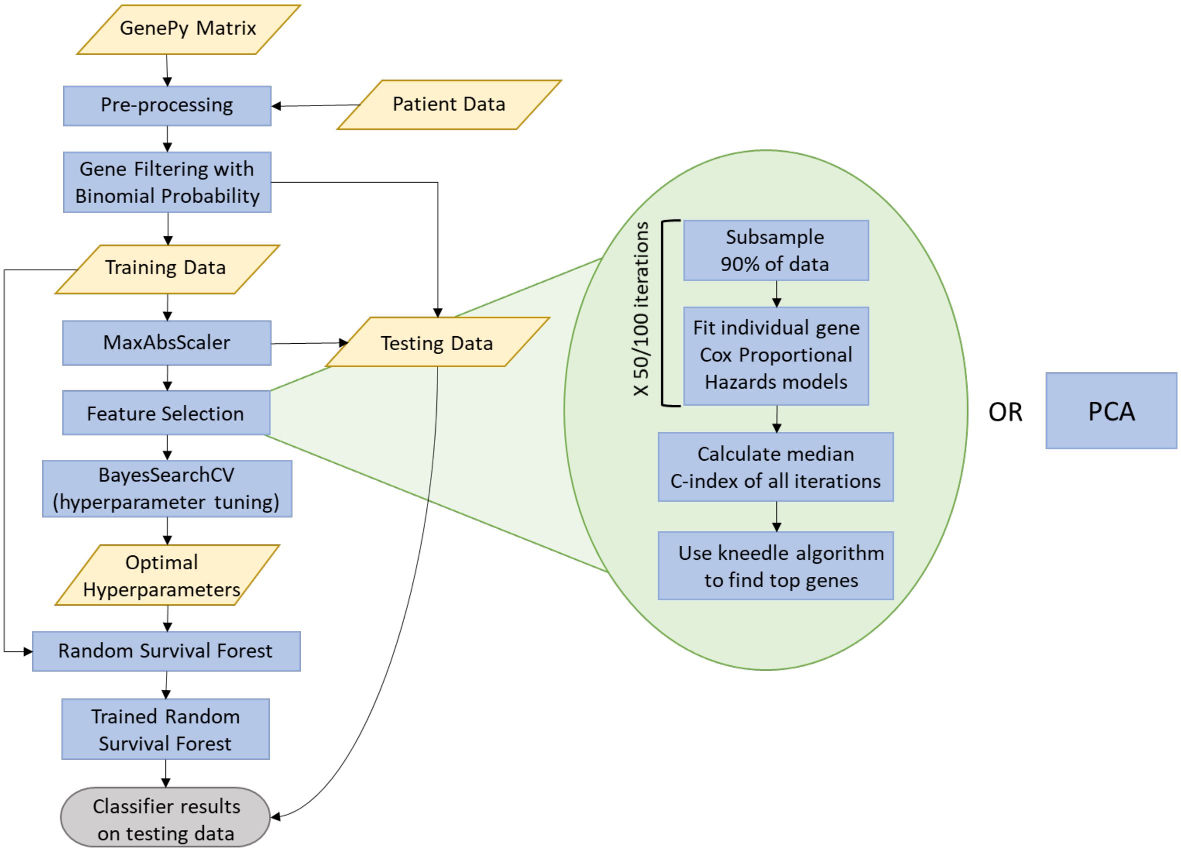

GenePy Matrix
Pre-processing
Patient Data
Gene Filtering with Binomial Probability
Training Data
MaxAbsScaler
Testing Data
Feature Selection
BayesSearchCV (hyperparameter tuning)
Optimal Hyperparameters
Random Survival Forest
Trained Random Survival Forest
Classifier results on testing data
X 50/100 iterations
Subsample 90% of data
Fit individual gene Cox Proportional Hazards models
Calculate median C-index of all iterations
Use kneedle algorithm to find top genes
OR
PCA

Figure 55 The final RSF machine learning pipeline. Added to the pipeline, in comparison to Figure 54, was hyperparameter tuning using Bayesian Optimisation, and two feature selection methods: CPH model gene ranking with subsampling and selection with the kneedle method, and dimensionality reduction with PCA.

7.3 Results

Of the 681 patients diagnosed with CD in the IBD cohort, 600 were of European ancestry with high probability, and unrelated to any individuals in the cohort. Patients were also required to have a confirmed stricturing or not-stricturing endotype, and a recorded clinical follow-up time. After excluding patients that were missing this information, 553 CD patients were included in the model (cohort characteristics detailed in Table 43). The number of patients in the training and testing datasets for RSF modelling is included in Table 44. Fewer patients have a stricturing outcome in the cohort. Therefore, the training set was manually balanced to include an equal number of patients with and without stricturing behaviour, such that both outcomes could be learnt from equally during training.

Table 43 Clinical characteristics of the CD cohort used in RSF modelling , split by paediatric and adult IBD diagnosis.

		Paediatric IBD (<18)	**Adult IBD (≥18)**	**Total**
N		314	239	553
Median age at diagnosis, years (range)		13 (1-17)	31 (18-82)	NA
Stricturing Endotype	Yes	68	95	163
	No	246	144	390
Sex	Male	195	102	297
	Female	119	137	256
Median follow-up time, years (range)		6 (0.1-54.8)	13 (2.4-49.8)	NA

Table 44 Training and Testing Data sample size by patient endotype

	Training Data	Testing Data	Total
Stricturing	122	41	163
Not-Stricturing	122	268	390
Total	244	309	553

7.3.1 Preliminary results: NOD-signalling pathway gene panel

The extended NOD-signalling pathway gene panel comprised 180 genes and was sourced from the Comparative Toxicogenomics Database, and combined the KEGG:hsa04621 and REACT:R-HSA-168638 signalling pathways (see Supplementary Files). After pre-processing, where invariant genes were removed, 130 genes remained. The gene set was reduced, such that there was a 95% probability that a random selection of individuals for the training set gave linearly independent columns (variables). This resulted in 44 genes as input into feature selection. Feature selection ranked these genes by their Cox Proportional Hazard (CPH) Model C-index (Table 45). The top 5 features were selected by BayesSearchCV, and these were fed in to the RSF model. On the training dataset the RSF achieved a C-index of 0.827, and on the testing dataset the C-index was 0.546. The average AUC for the training data was 0.876, and on testing set it was 0.559. The testing set AUC increased for patients with longer follow-up (Figure 56).

Table 45 Genes included in feature selection ranked by CPH model C-index

Gene	**CPH Model C-index**	**Gene**	**CPH Model C-index**	**Gene**	**CPH Model C-index**
CTSB	0.566788	TRPM7	0.521089	ITPR2	0.502453
PLCB3	0.561486	NLRP3	0.520772	IRAK2	0.501899
GBP3	0.554048	RIPK2	0.518557	NLRP12	0.499763
TRPV2	0.553889	ITPR1	0.517014	RIPK3	0.499011
MAPK12	0.551357	PLCB2	0.514521	P2RX7	0.498892
CYBA	0.550684	NOX1	0.513017	NOX3	0.498615
XIAP	0.539804	BIRC2	0.512978	CASR	0.496756
NOD2	0.539804	NLRX1	0.51274	CASP8	0.496637
ITPR3	0.533354	TNFAIP3	0.511118	RNF31	0.494302
IFI16	0.528488	GPRC6A	0.510881	SHARPIN	0.491849
CHUK	0.527222	MFN2	0.508982	CARD6	0.48362
TRPM2	0.526074	NOD1	0.507874	NFKBIB	0.480573
DHX33	0.525837	CASP5	0.507043	RNASEL	0.480454
TYK2	0.523859	GBP1	0.50641	ERBIN	0.477131
MAVS	0.522672	TP53BP1	0.506014		

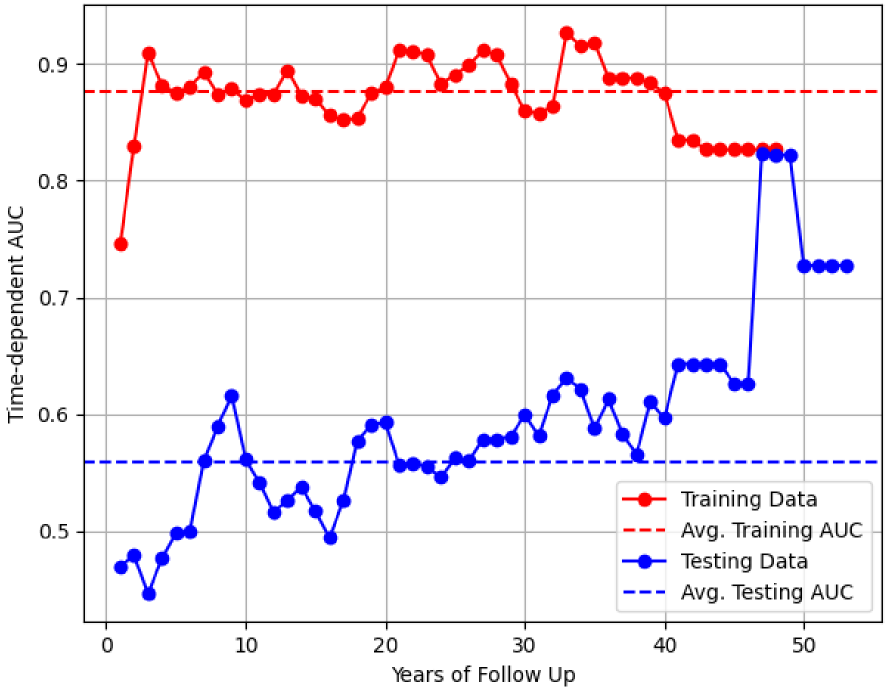


Figure 56 AUC over time for preliminary results on the NOD-signalling pathway gene panel. Average AUC for the RSF model on the training and testing data

Due to the relatively small sample sizes, particularly with respect to the number of stricturing endotype patients, there was a potential question surrounding how stable the results achieved with the NOD-signalling panel were. Therefore, the pipeline was run again, with a different initial seed that split the training and testing data, leading to a different set of patients being included in each dataset. The features selected, and subsequent performance of the RSF differed from the first set of results in Table 45 and Figure 56. With different patients in the training dataset, 15 genes were selected for inclusion in the RSF model, and of the features selected, only 2 of 5 were included in both models. The RSF model achieved a C-index of 0.957 on the training set, and 0.534 on the testing set (a 15% and 2% difference compared to previous results, respectively). The average AUC was 0.962 on the training set, and 0.509 on the testing set (a 9.8% difference for both training and testing AUCs). In light of these results, and in particular the discrepancies between the genes selected, the next stage was to look at how to stabilise the feature selection process.

7.3.2 Feature selection stability

In order to assess the feature selection stability, the 90% of the training data was subsampled randomly for a total of 50 trials. For each trial, the top number of features selected by BayesSearchCV (30 iterations) was recorded. While it would be impractical to exhaustively search larger gene panels, as these trials were performed on the NOD-signalling gene panel (n=44), the GridSearchCV method was also explored as a potential alternative to selecting the top number of features for modelling. This was to eliminate the possibility that variability in the number of genes chosen by the Bayes Search so far was due to the algorithm not finding the optimum number of features. The number of features selected varied widely for each trial (Figure 57), with some trials choosing only 2 or 3 features, while others selected all features. Furthermore, the ordering of the features by CPH model C-index in each trial was different. This meant that even if the same number of features was chosen over multiple trials, due to the different C-indices for each feature, different genes would be present in this top k features (see Supplementary files). Bayes Search and Grid Search were broadly concordant (36/50, 72%), which provided confidence that either method would be suitable for determining the top *k* features.

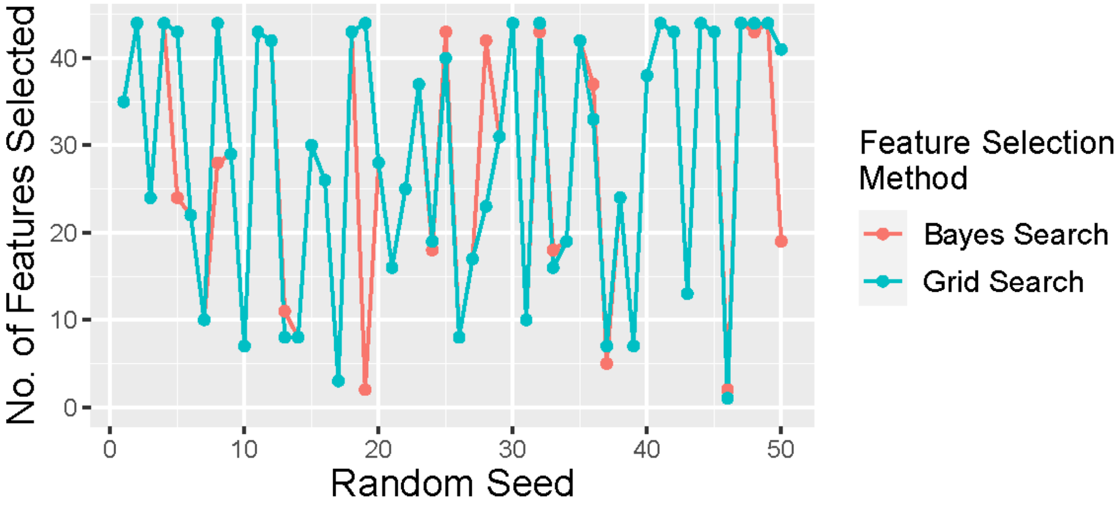


Figure 57 Number of features selected according to feature ranking in 50 trials. Two methods were used to choose the number of top features to input into the RSF model: Bayes Search and Grid Search. This selection was performed after ranking of genes by CPH model C-index. The random seed controls the samples that are including in the 90% subsample prior to gene ranking.

To mitigate the discovered variability in feature selection, the median value of the C-index of individual gene CPH models was tested. The median was used so that any C-index scores that were outliers would be less influential. As before, a random 90% of the training data was input into generating these CPH models. It was unknown how many iterations would be necessary before the C-index of each gene's model was stabilised, so the median score after 10, 20, 30, 40 and 50 iterations was graphed (Figure 58). Some genes exhibited very stable C-indices after only 10 iterations, such as *CYBA*, *ERBIN* and *PLCB2*. Others, such as *NOD2* and *P2RX7* required over 30 iterations of this feature selection process before the C-index scores became constant. After monitoring the changes in CPH C-index, taking the median after 50 iterations was judged to give a reasonably accurate representation of each gene's score, that was independent of the individual patients present in the training dataset.

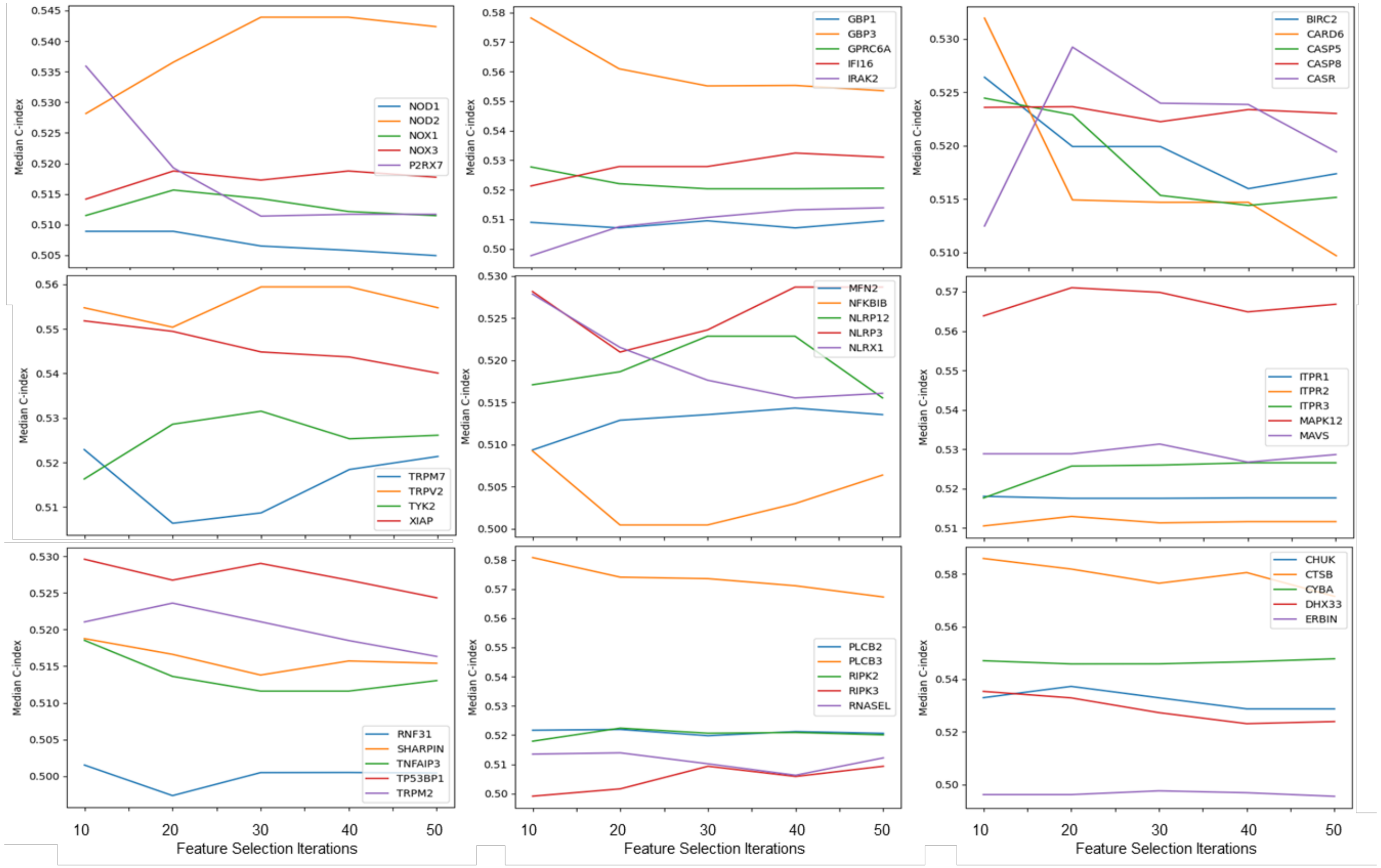


Figure 58 Median C-index of individual gene CPH models for 10, 20, 30, 40 and 50 iterations of feature selection on a random 90% of the training data (NOD-signalling panel, n=44). Genes were grouped alphabetically.

Once the feature rankings according to the median of 50 CPH model C-indices had been calculated, four experiments with different genes used in the RSF model were conducted to test if feature selection was actually impacting model performance: 1) the top 10 genes; 2) the bottom 10 genes; 3) 10 random genes; and 4) all 44 genes used. The best performing RSF was the one that utilised the top 10 genes, with a training and testing C-index of 0.959 and 0582, respectively, and similar values for average AUC (0.984 and 0.561). There were small margins between the other three models in terms of test C-indices and average AUCs over time. The next best RSF model used all 44 genes (C-index 0.538, average AUC 0.511 on the test dataset), followed by the model that used 10 random genes (test data C-index 0.518, average AUC 0.505), and lastly the model that used the bottom 10 genes (C-index 0.507, average AUC 0.502). This is the order of best performing to worst performing ML model that would be expected if the feature selection has a meaningful impact on the accuracy of the RSF. All models exhibit overfitting to some extent (Figure 59). The model that appears to be impacted the least by overfitting is the RSF where all 44 genes are included.

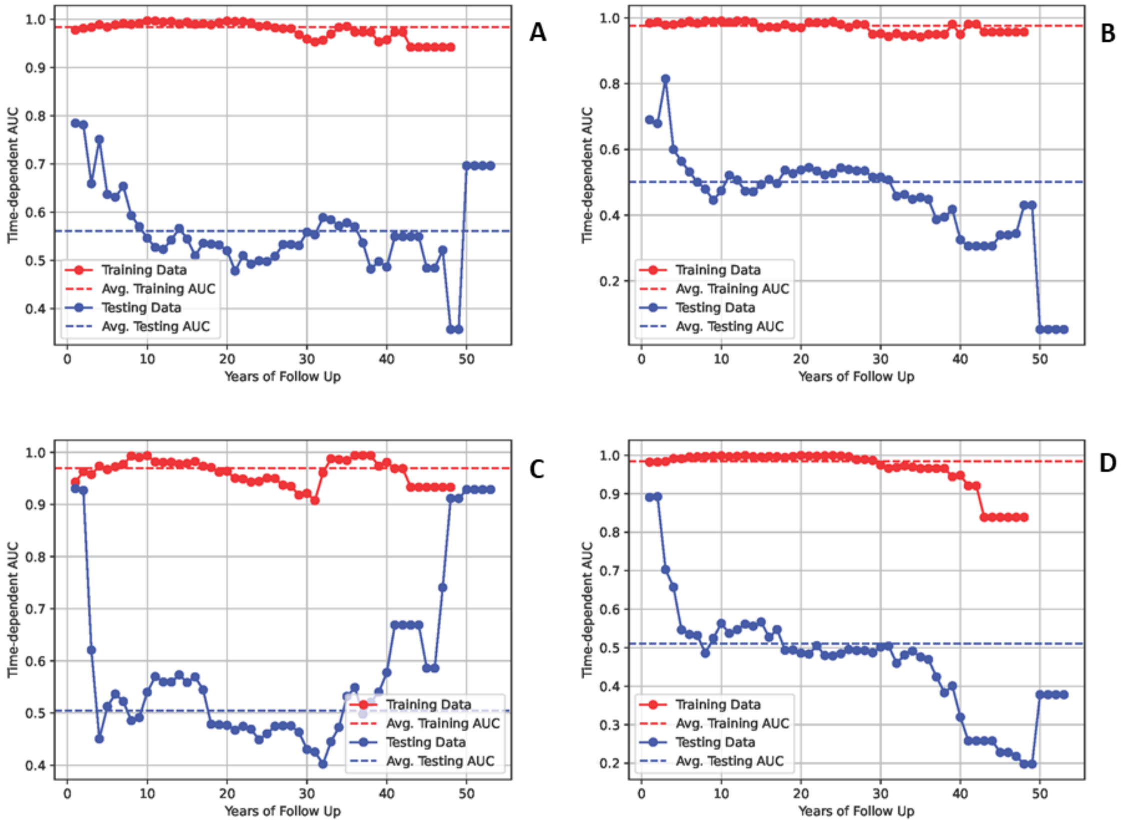


Figure 59 Random survival forest AUC over time for the training and testing dataset. A) top 10 features in the NOD-signalling panel, B) bottom 10 features in the NOD-signalling panel, C) 10 random genes from the NOD-signalling panel, D) all 44 features in the NOD-signalling panel

The feature weights for the RSF models using the top 10, bottom 10, and random 10 genes are listed in Table 46. For all models, the feature weights are small, with relatively little difference in their values when comparing the weights across the included genes. Further, many of the confidence intervals for each gene's weight have large confidence intervals, in some cases larger than the feature weight. It is difficult therefore, to determine with certainty whether each gene has either a protective effect against stricturing, or contributes to the phenotype. *NOD2* is notable for its inclusion in the top 10 features list, as this was one of the top 10 most important genes in several of the stricturing classifiers in Chapter 6, where different gene panels were trialled. It is surprising to see *P2RX7* in the bottom 10 features list, as this was also in the top 10 most important genes in several classifiers in Chapter 6.

Table 46 Feature weights for the RSF that includes the 1) top 10 features; 2) the bottom 10 features; and 3) 10 random features

Top 10 Features		**Bottom 10 Features**		**Random 10 Features**	
Gene	**Weight**	**Gene**	**Weight**	**Gene**	**Weight**
CHUK	0.0367 ± 0.0640	*NOD1*	0.0378 ± 0.0401	*SHARPIN*	0.0249 ± 0.0573
CTSB	0.0332 ± 0.0260	*GBP1*	0.0218 ± 0.0287	*DHX33*	0.0070 ± 0.0440
MAPK12	0.0307 ± 0.0407	*ERBIN*	0.0200 ± 0.0555	*NLRX1*	0.0061 ± 0.0211
GBP3	0.0304 ± 0.0816	*NFKBIB*	0.0141 ± 0.0375	*TNFAIP3*	0.0055 ± 0.0121
IFI16	0.0293 ± 0.0493	*CARD6*	0.0137 ± 0.0307	*NOD1*	0.0028 ± 0.0386
CYBA	0.0259 ± 0.0581	*NOX1*	0.0121 ± 0.0185	*CASR*	-0.0181 ± 0.0378
TRPV2	0.0187 ± 0.0346	*RNF31*	0.0121 ± 0.0185	*TRPM7*	-0.0186 ± 0.0259
NOD2	0.0120 ± 0.0353	*ITPR2*	0.0103 ± 0.0218	*RNASEL*	-0.0206 ± 0.0482
XIAP	0.0051 ± 0.0521	*P2RX7*	0.0031 ± 0.0687	*CASP8*	-0.0263 ± 0.0480
PLCB3	0.0009 ± 0.0306	*RIPK3*	-0.0104 ± 0.0220	*GPRC6A*	-0.0279 ± 0.0464

It had now been determined that this feature selection method of taking the median C-index of several iterations of CPH models, was successful for generating a better performing RSF model downstream. However, in the tests on the NOD-signalling pathway gene panel an arbitrary cut-off of 10 was used. The kneed package was used to find the knee of the ranked median CPH C-indices. After the knee point, the inclusion of additional genes is expected to result in minimal gains to RSF performance. The final pipeline for obtaining RSF model results on three different gene panels was therefore as follows (Figure 55, Section 7.2.2):

- Initial gene panel reduced to a set of genes where every column (gene) is linearly independent. This was done with binomial probability.
- Feature selection on a random 90% of the dataset for 50 iterations (NOD-signalling gene panel) or 100 iterations (stricturing (inclusive) gene panel, and IBD gene panel, as these were larger panels). Each feature iteration produced a CPH model C-index for each gene.
- The knee of the median C-index after 50 or 100 iterations (depending on panel) was found.
- Hyperparameter tuning using Bayes Search in a nested cross-validation scheme was performed, using the selected features (hyperparameter values, Table 47).
- Optimal hyperparameter values were used in the RSF, along with the selected features, and the model performance assessed on the testing dataset.

Table 47 Hyperparameter default values, and values tested during individual, grid search and Bayes hyperparameter tuning for the stricturing endotype classifier

Hyperparameter Name	**Default**	**Values Included for Bayes Hyperparameter Tuning**
n_estimators	100	100, 250, 500, 750, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 9000, 10000
max_depth	None	1 - 30, None
min_samples_split	2	2, 3, 4, 5, 6
min_samples_leaf	1	1, 2, 3, 4, 5, 6
max_features	sqrt	sqrt(n_features), log2(n_features), None

7.3.3 NOD-signalling panel random survival forest

As previously discussed, the NOD-signalling pathway’s gene panel was reduced from 180 genes to 44 genes after all GenePy score matrix data processing steps. The knee of the median C-index scores after 50 iterations of feature selections was found to be at 9, and therefore 10 genes were chosen as input into the RSF (python indices start at 0), shown in Figure 60. This was coincidentally the same number of genes that was arbitrarily chosen in the previous section.

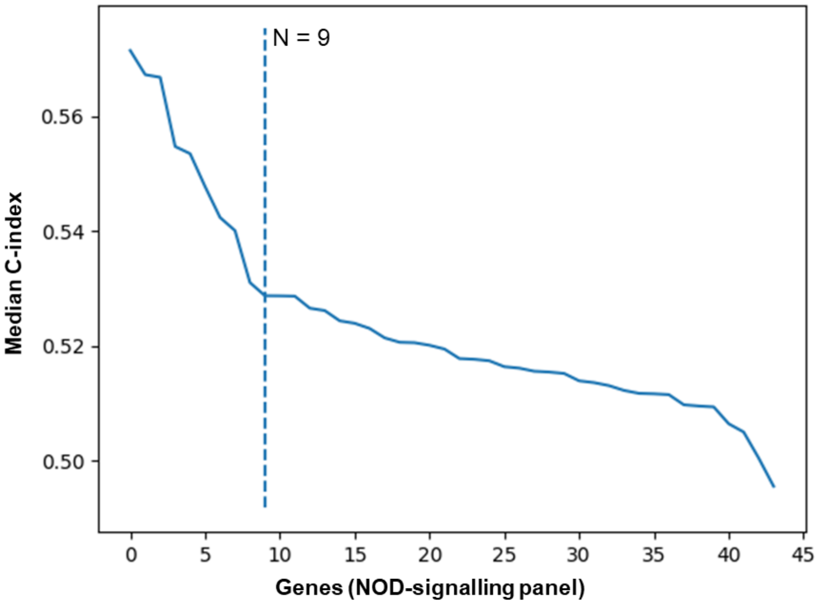


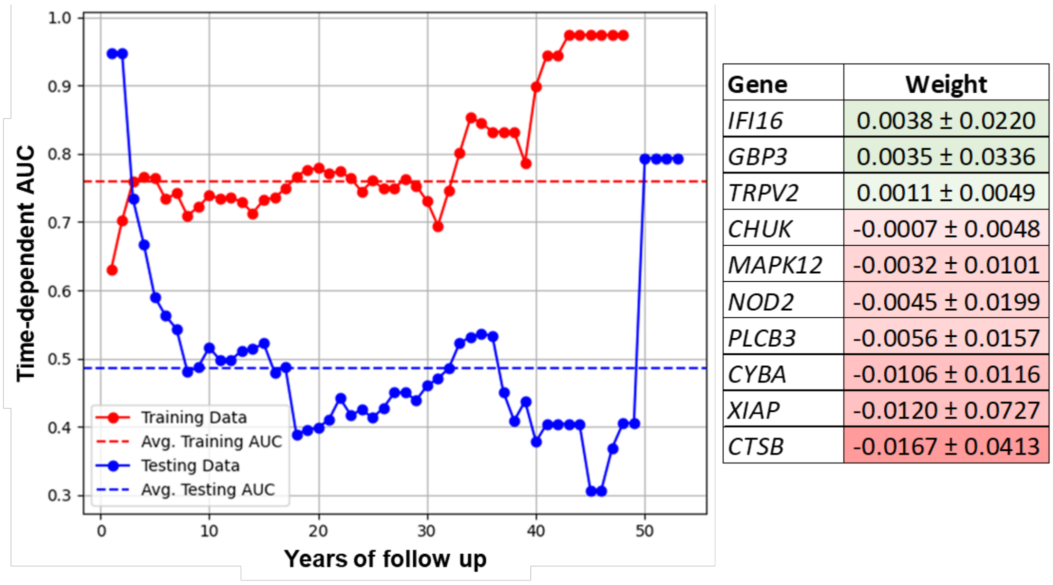
Figure 60 Median C-index for each gene in the NOD-signalling gene panel after 50 iterations of feature selection. Dotted line shows the knee in the dataset.

Bayes Search hyperparameter tuning (60 iterations) was performed in a nested cross validation scheme, with 3 folds in the inner cross validation, and 5 folds in the outer. Each outer fold had an equal number of stricturing and not-stricturing patients. Table 48 contains the results of the hyperparameter tuning. The hyperparameter values for maximum depth were small (1, 2, or 3), and the hyperparameter values for the minimum samples per split were relatively large across folds. The average C-index across the outer folds was 0.587 (standard deviation 0.072). The best C-index in the outer fold was 0.669 in fold 5. The chosen hyperparameters for the final RSF model were therefore: maximum depth = 2, maximum number of features = log2, minimum samples per leaf = 1, minimum samples per split = 6, and 750 estimators.

Table 48 BayesSearchCV hyperparameter tuning (CV folds, 1-3 n=48; 4 and 5 n=50), on the NOD-signalling gene panel after feature selection (n=10). Nested scheme with 3-fold inner cross validation (CV), 5-fold outer CV, 60 iterations.

Fold	Outer Fold C-Index	Inner CV C-index	Optimal Hyperparameters				
			Max Depth	Max Features	Minimum Samples per Leaf	Minimum Samples per Split	Number of Estimators
1	0.6311	0.5411	1	Sqrt	1	6	500
2	0.62107	0.6007	1	Sqrt	1	5	3000
3	0.4660	0.6266	3	None	5	5	4000
4	0.5492	0.6156	1	Log2	5	4	100
5	0.6687	0.5527	2	Log2	1	6	750

The performance of the RSF model using the NOD-signalling gene panel, and the corresponding feature weights can be viewed in Figure 61. This model performs less well than the one in Section 7.3.2, without hyperparameter tuning, with a C-index of 0.682 on the training data, and 0.514 on the testing data. However, in comparison to Figure 59A, this model's time-dependent AUC shows less overfitting on the training data. These results may therefore be a more accurate reflection of how an RSF model will perform on this data, and the model constructed will be more generalisable. The testing data's AUC for patients with relatively little clinical follow up data is very high (AUC > 0.9), perhaps indicating the model has detected some genomic signal from patients who stricture relatively early in their disease course. As in previous results, the weights of each gene are small, and have very large confidence intervals. Many genes with positive weights in Table 46, now have negative weights in this model, caused by the change in hyperparameters.



Gene	Weight
IFI16	0.0038 ± 0.0220
GBP3	0.0035 ± 0.0336
TRPV2	0.0011 ± 0.0049
CHUK	-0.0007 ± 0.0048
MAPK12	-0.0032 ± 0.0101
NOD2	-0.0045 ± 0.0199
PLCB3	-0.0056 ± 0.0157
CYBA	-0.0106 ± 0.0116
XIAP	-0.0120 ± 0.0727
CTSB	-0.0167 ± 0.0413

Figure 61 AUC over time of the RSF model on the training and testing datasets, with the feature weights of each gene included in the model (NOD-signalling gene panel).

7.3.4 Stricturing panel random survival forest

This panel, referred to as the stricturing (inclusive) gene panel, was selected to use as an input as it includes three gene panels: 1) the autoimmune gene panel; 2) the IBD gene panel; and 3) the stricturing gene panel from the literature search in Chapter 6. In total, this panel contains 3,155 genes, which following pre-processing steps of removing false positive, and invariant genes, left 2,192 genes. After processing the data so all columns (gene's GenePy scores) were linearly independent, there were 666 genes for input into feature selection. As this gene panel was the largest used in this modelling pipeline, the dataset was naturally sparser in comparison. Therefore, the binomial probability for producing a linearly independent dataset had to be increased to 99%, in order for downstream nested cross validation during hyperparameter tuning to be performed. The knee of the median C-indices after 100 subsamples was found to be at 38, shown in Figure 62, therefore 39 genes were included in downstream modelling (python indices start at 0).

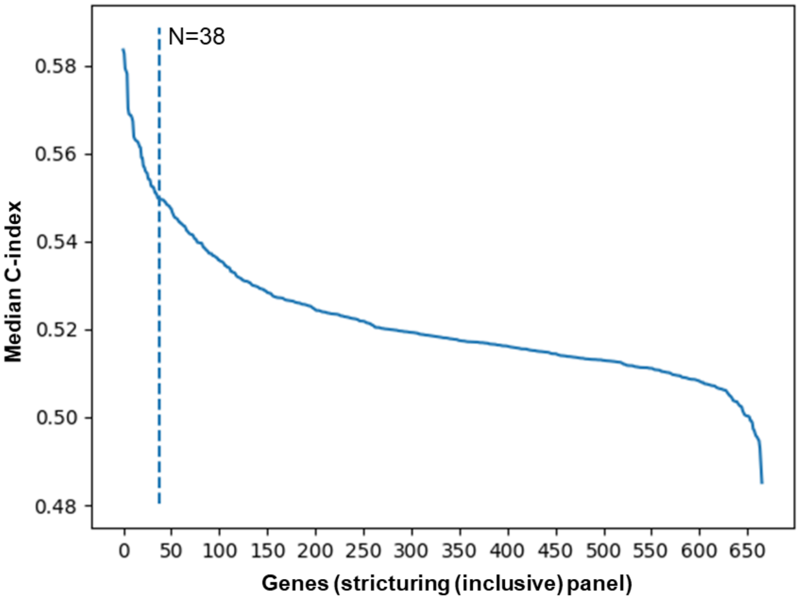


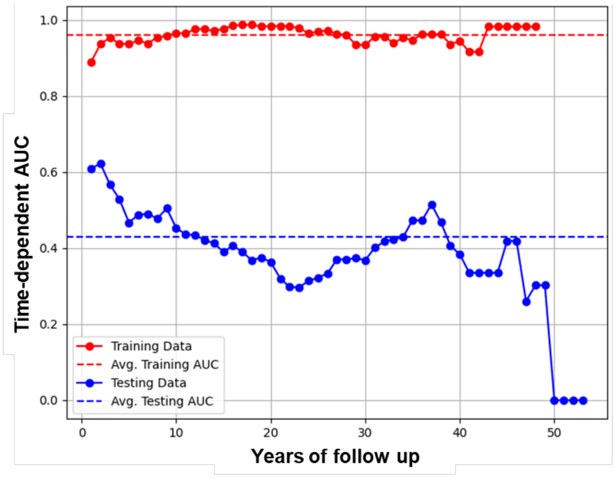
Figure 62 Median C-index for each gene in the stricturing (inclusive) gene panel after 100 iterations of feature selection. Dotted line shows the knee in the dataset.

Bayes Search hyperparameter tuning (60 iterations) was performed in the same nested cross validation scheme as for NOD-signalling gene panel. The stricturing and not-stricturing classes were balanced in each fold. The results of the hyperparameter tuning can be viewed in Table 49. The hyperparameter value for the number of estimators selected was either a few hundred trees, or several thousand. Number of samples required per leaf and split were relatively high, and max features was small (log2(39 genes) ≈ 5). The average C-index across the outer folds was 0.687 (standard deviation 0.037). The best C-index in the outer fold was 0.736 in fold 2. The chosen hyperparameters for the final RSF model were therefore: maximum depth = 7, maximum number of features = log2, minimum samples per leaf = 4, minimum samples per split = 3, and 4000 estimators.

Table 49 BayesSearchCV hyperparameter tuning (CV folds, 1-3 n=48; 4 and 5 n=50), on the stricturing (inclusive) gene panel after feature selection (n=39). Nested scheme with 3-fold inner cross validation (CV), 5-fold outer CV, 60 iterations.

Fold	**Outer Fold C-Index**	**Inner CV C-index**	**Optimal Hyperparameters**				
			Max Depth	**Max Features**	**Minimum Samples per Leaf**	**Minimum Samples per Split**	**Number of Estimators**
1	0.7016	0.6894	29	Log2	1	6	250
2	0.7357	0.6651	7	Log2	4	3	4000
3	0.6577	0.6978	6	Log2	2	5	6000
4	0.6331	0.7243	23	Sqrt2	2	3	250
5	0.7053	0.6802	19	Log2	3	6	250

The RSF model performance, and corresponding feature weights for the stricturing (inclusive) gene panel is shown in Figure 63. This ML model had a C-index of 0.914 on the training data, and 0.457 on the testing data. The C-index and average AUC is affected by the model's poor prediction for patients with over 45 years of clinical follow up. As in the NOD-signalling gene panel model, there is an uptick in performance for patients with very little follow up (AUC≈0.6), although to a lesser extent. This model shows clear signs of overfitting on the training data, with an AUC continuously very close to 1.0. The gene *COL6A2* has one of the largest weights seen in RSF modelling so far. It also has a confidence interval which confirms that a high GenePy score (high pathogenicity burden) in this gene would always be expected to contribute to stratification of patients into the stricturing group.



Gene	Weight	Gene	Weight
COL6A2	0.0428 ± 0.0329	*MAP9*	-0.0014 ± 0.0260
ANKRD12	0.0202 ± 0.0195	*PLCL1*	-0.0015 ± 0.0120
GSTO1	0.0103 ± 0.0083	*PLCB3*	-0.0024 ± 0.0095
PROCR	0.0092 ± 0.0097	*ELP1*	-0.0026 ± 0.0170
TTC7A	0.0089 ± 0.0098	*LOXL1*	-0.0032 ± 0.0277
NQO1	0.0088 ± 0.0099	*PODNL1*	-0.0035 ± 0.0146
LRR1	0.0054 ± 0.0085	*CYP1B1*	-0.0053 ± 0.0185
C1QTNF6	0.0040 ± 0.0121	*TRIM22*	-0.0057 ± 0.0082
RGS12	0.0040 ± 0.0118	*PLEK*	-0.0059 ± 0.0117
GBP3	0.0030 ± 0.0198	*NRDE2*	-0.0078 ± 0.0087
TRPV2	0.0029 ± 0.0055	*SLC39A11*	-0.0078 ± 0.0188
P2RX4	0.0025 ± 0.0122	*CNTRL*	-0.0092 ± 0.0183
MAPK12	0.0023 ± 0.0080	*MFSD9*	-0.0100 ± 0.0167
ICAM2	0.0020 ± 0.0096	*CDKN2D*	-0.0103 ± 0.0180
FCER2	0.0013 ± 0.0101	*HIT*	-0.0116 ± 0.0195
USP18	0.0009 ± 0.0129	*FKBP10*	-0.0116 ± 0.0194
GALC	0.0009 ± 0.0278	*ADAMTS3*	-0.0125 ± 0.0207
CTSB	0.0002 ± 0.0234	*OTUD7B*	-0.0132 ± 0.0488
COL9A1	-0.0003 ± 0.0097	*TNFSF13*	-0.0145 ± 0.0169
MMP10	-0.0007 ± 0.0107		

Figure 63 AUC over time of the RSF model on the training and testing datasets, with the feature weights of each gene included in the model (stricturing (inclusive) gene panel).

7.3.5 IBD panel random survival forest

This panel was selected, as in Section 6.3.2 it was shown to deliver the best performing random forest classifier for stricturing endotype classification. This panel contains 821 genes, which following pre-processing steps of removing false positive, and invariant genes, left 465 genes. After filtering the data so all columns (gene's GenePy scores) were linearly independent, there were 159 genes for input into feature selection. The knee of the median C-indices after 100 subsamples was found to be at 26, shown in Figure 64, therefore 27 genes were included in downstream modelling (python indices start at 0).

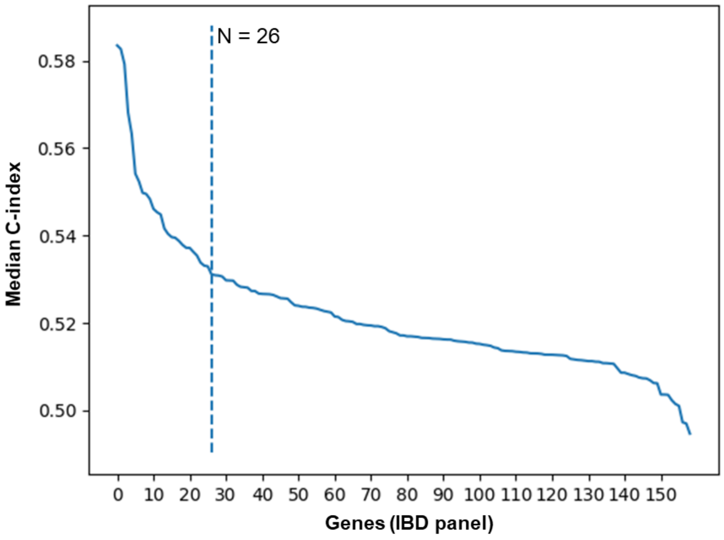


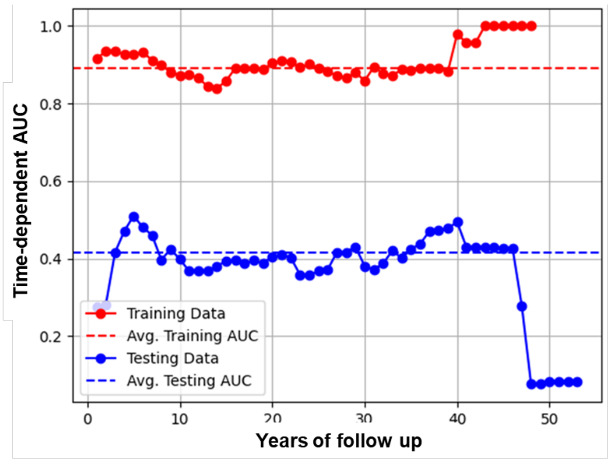
Figure 64 Median C-index for each gene in the IBD gene panel after 100 iterations of feature selection. Dotted line shows the knee in the dataset.

Bayes Search hyperparameter tuning (60 iterations) was performed in the same nested cross validation scheme as the previous two models. The stricturing and not-stricturing classes were balanced in each fold. The results of the hyperparameter tuning can be viewed in Table 50. The hyperparameter value for the number of estimators selected was consistently large. The number of samples required per split was consistently relatively high. The maximum depth was small across all folds, as was the maximum features per split (log2(27 genes) ≈ 5). The average C-index across the outer folds was 0.660 (standard deviation 0.053). The best C-index in the outer fold was 0.736 in fold 5. The chosen hyperparameters for the final RSF model were therefore: maximum depth = 4, maximum number of features = log2, minimum samples per leaf = 4, minimum samples per split = 4, and 8000 estimators.

Table 50 BayesSearchCV hyperparameter tuning (CV folds, 1-3 n=48; 4 and 5 n=50), on the IBD gene panel after feature selection (n=27). Nested scheme with 3-fold inner cross validation (CV), 5-fold outer CV, 60 iterations.

Fold	Outer Fold C-Index	Inner CV C-index	Optimal Hyperparameters				
			Max Depth	Max Features	Minimum Samples per Leaf	Minimum Samples per Split	Number of Estimators
1	0.6094	0.6823	1	Log2	1	5	4000
2	0.7006	0.6226	3	Log2	1	4	1000
3	0.5959	0.6435	1	Log2	1	4	9000
4	0.6595	0.6504	2	Log2	1	5	9000
5	0.7358	0.6049	4	Log2	4	4	8000

RSF time-dependent AUC and input gene weights are shown in Figure 65. This model does show less overfitting on the training dataset than the stricturing (inclusive) panel model. Unfortunately, a relatively more generalisable model does not translate to better testing performances, with a poor test set C-index, at 0.438 (training dataset C-index 0.826). The testing AUC is poor (less than 0.5) regardless of the number of years of clinical follow up. Despite the IBD gene panel generating best performing classifier for the stricturing endotype in Section 6.3.2, of the three gene panels tested with RSF, this panel produced the poorest model. The trends of small gene weights and large confidence intervals are present here as in the previous two models. The two genes with the largest weights have small enough confidence intervals that their effects can be confirmed as positive or negative. *NOD2* has the largest positive weight towards stricturing, and *TRIM22* has the largest weight against the stricturing endotype.



Gene	Weight	Gene	Weight
NOD2	0.0231 ± 0.0149	*SLAMF8*	-0.0007 ± 0.0030
TTC7A	0.0207 ± 0.0212	*XIAP*	-0.0010 ± 0.0147
PROCR	0.0093 ± 0.0193	*TMEM17*	-0.0021 ± 0.0083
UTP20	0.0077 ± 0.0098	*FAM118A*	-0.0027 ± 0.0065
DENND1B	0.0065 ± 0.0106	*MTMR3*	-0.0028 ± 0.0108
SLC22A4	0.0055 ± 0.0033	*TET2*	-0.0029 ± 0.0147
SPATA48	0.0045 ± 0.0037	*CYBA*	-0.0040 ± 0.0068
GPR65	0.0039 ± 0.0085	*SLC39A11*	-0.0066 ± 0.0269
CELSR3	0.0037 ± 0.0206	*LRRK2*	-0.0106 ± 0.0135
MUS81	0.0029 ± 0.0187	*CNTRL*	-0.0123 ± 0.0214
GSDMB	0.0012 ± 0.0044	*MFSD9*	-0.0165 ± 0.0187
PLCB3	0.0009 ± 0.0089	*GALC*	-0.0245 ± 0.0849
CCL8	0.0006 ± 0.0040	*TRIM22*	-0.0250 ± 0.0195
PLCL1	0.0003 ± 0.0139		

Figure 65 AUC over time of the RSF model on the training and testing datasets, with the feature weights of each gene included in the model (IBD gene panel).

7.3.6 *NOD2* only random survival forest

Due to recent survival modelling results that used only *NOD2* GenePy scores to stratify patients into multiple stricturing endotype risk groups (unpublished data), there was interest in how an RSF model would perform if the only feature present was *NOD2.* Therefore the RSF was hyperparameter tuned, as in previous sections, with only *NOD2* as the input. The results of this hyperparameter tuning can be viewed in Table 51. All hyperparameters took on many values, with no strong trends across cross validation folds. Of all models hyperparameter tuned in this chapter, this model had the lowest average C-index across folds (0.506, standard deviation 0.064). The best C-index in the outer fold was 0.610 in fold 1. The chosen hyperparameters for the final RSF model were therefore: maximum depth = 5, minimum samples per leaf = 2, minimum samples per split = 5, and 250 estimators.

Table 51 BayesSearchCV hyperparameter tuning (CV folds, 1-3 n=48; 4 and 5 n=50), when including only NOD2. Nested scheme with 3-fold inner cross validation (CV), 5-fold outer CV, 60 iterations.

Fold	**Outer Fold C-Index**	**Inner CV C-index**	**Optimal Hyperparameters**			
			Max Depth	**Minimum Samples per Leaf**	**Minimum Samples per Split**	**Number of Estimators**
1	0.6085	0.5484	5	2	5	250
2	0.4815	0.5904	2	1	3	4000
3	0.4598	0.5807	4	4	4	250
4	0.5480	0.5105	2	5	6	100
5	0.4319	0.5483	7	2	3	100

The RSF model utilising *NOD2* only achieves a C-index of 0.623 on the training dataset, and 0.454 on the testing dataset (AUC over time, Figure 66). As expected, there is less overfitting present, due to there being only one feature. Model prediction is particularly poor for patients with less than 5 years of clinical follow up, and patients with over 46 years of follow up. This model's performance is on a par with the models generated using the stricturing (inclusive) panel, and the IBD gene panel, but is outperformed by the NOD-signalling pathway gene panel.

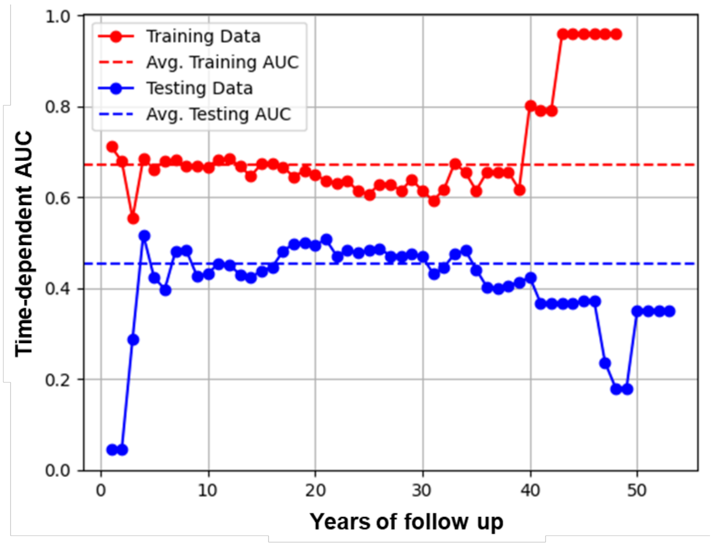


Figure 66 AUC over time for RSF model on the training and testing datasets, with only NOD2 as a feature.

7.3.7 PCA as an alternative to CPH model feature selection

The use of CPH models as a feature selection method was advantageous for this type of longitudinal time-to-stricture modelling. However, it was also thought that a method for dimensionality reduction, PCA, may be better equipped to deal with the sparse genomic data. Four different feature selection experiments were conducted. These combined PCA with other methods that had been used in the previous pipeline. These experiments and the results on the RSF model are summarised in Table 52.

Table 52 The testing and training C-indices from the RSF model for the four feature selection experiments (NOD-signalling gene panel)

Feature Selection Method	**Training Dataset C-index**	**Testing Dataset C-index**
PCA followed up selection of top PCs using knee method	0.965	0.464
PCA only	0.961	0.498
Dataset reduction using binomial probability followed by PCA	0.963	0.595
Dataset reduction using binomial probability followed by PCA, then knee method PC selection.	0.967	0.535

The NOD-signalling pathway genes were pre-processed as before (false positive, and invariant genes removed), leaving 130 genes to be reduced into principal components (PCs). These genes were transformed into 73 PCs that explained 95% of the variance in the data. In the first experiment, the knee of these PCs was located using the kneed package as before. This method selected the top 12 PCs for inclusion into further modelling. These chosen PCs accounted for 49.2% of the variance in the data. With these PCs as input for the RSF model, the testing data C-index was 0.464. One of the potential causes of this performance was that each PC explained only a fraction of the data; the first PC only accounted for 8.1% of the data variance. Therefore, for the second experiment all 73 PCs were included in the RSF model. This performed better than the first model, but still poorly (testing data C-index 0.498).

While with the implementation of PCA for dimensionality reduction, sub-setting the data to contain linearly independent columns was not necessary, this technique was trialled to see how it would affect the results. Hence, 44 linearly independent genes were transformed into 34 PCs that represented 95% of the variance. Every PC was input into an RSF model, generating the best performing model in this third experiment (test C-index 0.595). Finally, the knee of the 34 PCs was found, as in the first experiment. This resulted in the first 9 PCs (explains 59.0% of the variance) being input into the RSF model. The performance of the model subsequently decreased, with a test C-index of 0.535. It is worth noting that the RSF models generated in all four experiments showed clear signs of overfitting.

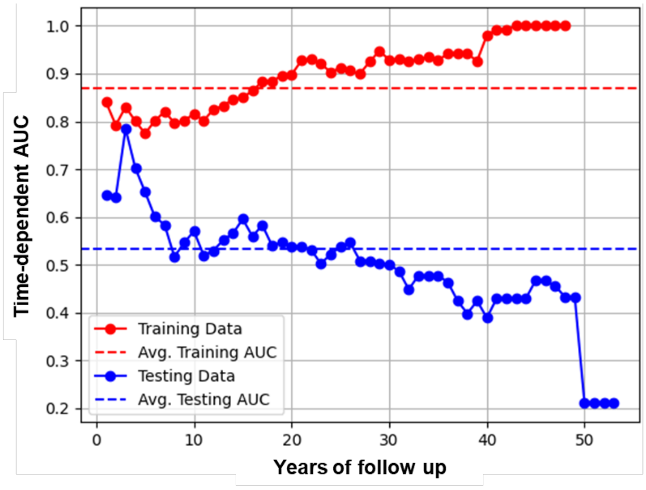
7.3.7.1 Final model: NOD-signalling gene panel, and dimensionality reduction with principal component analysis

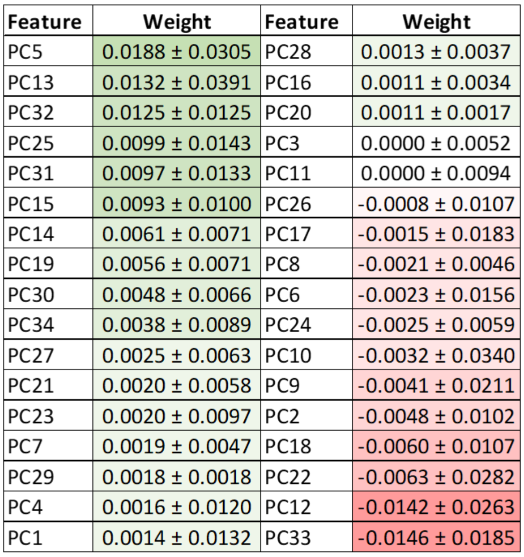
To find if the best performing model from the previous experiments (dataset reduction to a linearly independent subset, followed by PCA) could be improved, hyperparameter tuning was performed with this feature set, before generating a final RSF model. Bayes Search optimisation with 60 iterations was used, with the same cross validation scheme as all previous modelling. The results of the hyperparameter tuning can be viewed in Table 53. A wide range of estimator values were chosen across folds, from 250, to 9000. The values chosen for maximum depth were very small. The average C-index across the outer folds was 0.520 (standard deviation 0.075). The best C-index in the outer fold was 0.610 in fold 4. The chosen hyperparameters for the final RSF model were therefore: maximum depth = 2, maximum number of features = sqrt, minimum samples per leaf = 6, minimum samples per split = 2, and 5000 estimators. This combination of hyperparameters means that the required minimum number of samples per leaf will override the selected hyperparameter value for the minimum number of samples per split.

Table 53 BayesSearchCV hyperparameter tuning (CV folds, 1-3 n=48; 4 and 5 n=50), using 34 principal components. Nested scheme with 3-fold inner cross validation (CV), 5-fold outer CV, 60 iterations.

Fold	Outer Fold C-Index	Inner CV C-index	Optimal Hyperparameters				
			Max Depth	Max Features	Minimum Samples per Leaf	Minimum Samples per Split	Number of Estimators
1	0.4684	0.5328	1	sqrt	1	3	1000
2	0.5527	0.5279	1	None	4	4	750
3	0.4000	0.5848	3	Sqrt	1	6	250
4	0.6091	0.5214	2	Sqrt	6	2	5000
5	0.5691	0.5194	2	None	1	5	9000

The RSF model achieves a C-index of 0.774 on the training data, and a C-index of 0.576 on the test dataset (AUC over time, and model feature weights shown in Figure 67). Although the tuned model sees a slight drop in performance on the test dataset, the training data AUC suggests this model has less signs of overfitting than the untuned model. Therefore, this model is the best RSF in this chapter, as it combines generalisability with a good test dataset C-index. There are small features weights with large confidence intervals as in all previous modelling. PC5 and PC13 have the largest positive weights, while PC12 and PC33 have the largest negative weights. PC5 is characterised by a positive *SHARPIN* loading, and negative loadings from *CASP5*, *GPRC6A*, and *RNASEL*. PC13 has many comparatively large loadings in positive and negative directions: *CTSB*, *GBP1*, *ITPR1*, *ITPR2* and *RNF31* are all positive loadings, and *GBP3*, *ITPR3* and *PLCB2* provide negative loadings. In PC12, *BIRC2* and *TRPM2* are positive loadings, while *CTSB* and *P2RX7* give negative loadings. The strongest positive loading in PC33 is *NOX3*, with a value double the next-largest loading. *NOD2* and *MFN2* are the largest negative loadings in PC33. Of all the genes mentioned as present in the PCs with the largest weightings, only *CTSB*, *NOD2* and *GBP3* were present in the previous modelling that utilised the NOD-signalling gene panel (Section 7.3.3).





Feature	Weight	Feature	Weight
PC5	0.0188 ± 0.0305	PC28	0.0013 ± 0.0037
PC13	0.0132 ± 0.0391	PC16	0.0011 ± 0.0034
PC32	0.0125 ± 0.0125	PC20	0.0011 ± 0.0017
PC25	0.0099 ± 0.0143	PC3	0.0000 ± 0.0052
PC31	0.0097 ± 0.0133	PC11	0.0000 ± 0.0094
PC15	0.0093 ± 0.0100	PC26	-0.0008 ± 0.0107
PC14	0.0061 ± 0.0071	PC17	-0.0015 ± 0.0183
PC19	0.0056 ± 0.0071	PC8	-0.0021 ± 0.0046
PC30	0.0048 ± 0.0066	PC6	-0.0023 ± 0.0156
PC34	0.0038 ± 0.0089	PC24	-0.0025 ± 0.0059
PC27	0.0025 ± 0.0063	PC10	-0.0032 ± 0.0340
PC21	0.0020 ± 0.0058	PC9	-0.0041 ± 0.0211
PC23	0.0020 ± 0.0097	PC2	-0.0048 ± 0.0102
PC7	0.0019 ± 0.0047	PC18	-0.0060 ± 0.0107
PC29	0.0018 ± 0.0018	PC22	-0.0063 ± 0.0282
PC4	0.0016 ± 0.0120	PC12	-0.0142 ± 0.0263
PC1	0.0014 ± 0.0132	PC33	-0.0146 ± 0.0185

Figure 67 AUC over time of the RSF model on the training and testing datasets, with the feature weights of each principal component included in the model, based on the NOD-signalling pathway gene panel.

7.4 Discussion

In this chapter, five RSF models were fully tuned and trained for stratifying patients by stricturing endotype (Figure 68). Three gene panels were employed in RSF modelling: the NOD-signalling pathway gene panel, the IBD gene panel, and the stricturing (inclusive) gene panel. Gene panels were chosen for both biological and computation reasons. The NOD-signalling pathway gene panel was deemed as a good panel for initial testing, firstly for *NOD2*'s suggested significance to the development of stricturing disease [5, 388], and secondly as the number of genes included in the panel was relatively small (n=180). For all RSF modelling there was a concern regarding highly dimensional data (number of genes being greater than the number of features) affecting the ability of an RSF to correctly predict a stricturing endotype. This could be of particular concern here as the dataset size was the smallest in comparison to the stricturing and subtype classifiers of previous Chapters, and the RSF approach is a form of regression problem, which is naturally more complex than a binary classifier. The IBD gene panel was chosen in modelling as this achieved the best classification in Section 6.3.2, providing a useful comparison. Finally, the stricturing (inclusive) panel was used for modelling because, even though there could be some concerns about highly dimensional data (n=3,155), there was an interest in how the RSF model would perform with a gene panel that specifically considered the stricturing endotype, rather than IBD more generally. The construction and limitations of this gene panel have been previously discussed in Section 6.4.

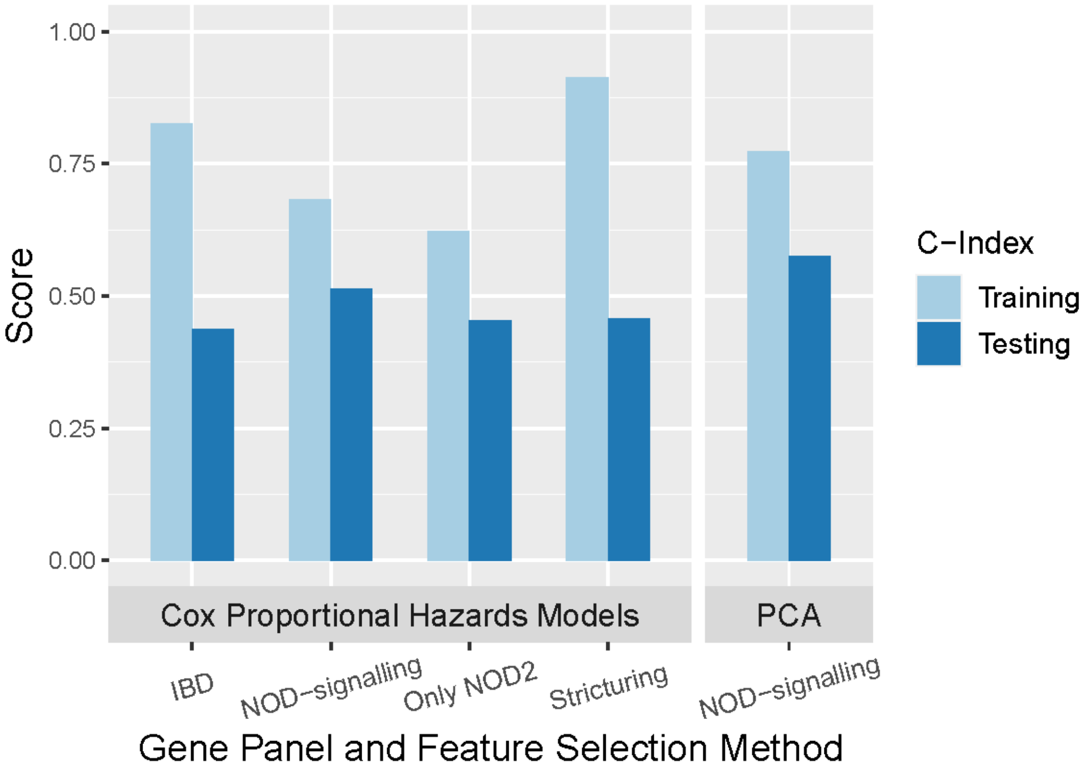


Figure 68 Summary of the five trained and tuned RSF model results. C-index score for training and testing for each of the gene panels used, for the two different feature selection approaches.

Three of the RSF models used CPH models during feature selection, starting with one of three different gene panels (NOD-signalling, IBD and stricturing (inclusive) panels), the fourth only used *NOD2* as a feature, and the fifth used the NOD-signalling pathway gene panel and PCA. This straightforward dimensionality reduction technique was more successful than the use of CPH models, which were originally chosen so that time-to-event data could be incorporated into the feature selection, as well as the machine learning method. Feature selection was shown to be beneficial, even with the CPH models, after experiments with the NOD-signalling panel revealed a better performing RSF using the top 10 genes, in comparison to all genes, a random 10 genes, and the bottom 10 genes. In addition, after modelling using only *NOD2*, which had been previously shown to be able to stratify stricturing patients, it was confirmed that including many genes in RSF models was beneficial, resulting in an increases in the testing C-index. Approaching the machine learning problem in this way was always going to be challenging, as a survival regression algorithm has to make predictions on a continuous scale, rather than a binary classification. Nevertheless, a modestly good model that achieved a testing C-index of 0.58 was generated using the NOD-signalling gene panel and PCA.

A new method for GenePy score matrix processing was employed for this chapter: reducing the data to a set of linearly independent columns (genes). It could be the case that through this method a gene is removed, in which a patient has a rare, highly scoring variant, and this results in a causal genomic signal being lost. However, the aim of the machine learning here is not to diagnose individuals, but rather to try and detect patterns present within groups in the cohort. Therefore, the trade-off that a few potential causal genes for single patients could be removed was seen as an acceptable loss in order to use the feature selection. In later modelling using PCA, the reduction in data sparsity that occurred using this method also proved beneficial for increasing the C-index and average AUC on the test dataset.

Unfortunately, the initial feature selection method of individual gene CPH models became a much more complex procedure than originally intended, with the many iterations that were required to establish a stable C-index for each gene. The necessity of this was discovered from initial modelling with the NOD-signalling gene pathway panel. When a different random seed was utilised, resulting in different individuals being included in the training data, the result was a different model in terms of the number of genes selected, which genes were selected for inclusion into RSF modelling, and a different testing AUC and C-index. This illustrated a key characteristic of the dataset, that the GenePy score distributions varied according to the patients included in the training data. This highlighted a limitation of the dataset, that any selected training dataset could not be relied upon to be representative enough of the population as a whole. Establishing a stable C-index through iterating CPH modelling on a subset of the training dataset attempted to increase the generalisability of the modelling, but it cannot be assumed that this wholly mitigated the genomic variability in the dataset.

In contrast, a simple PCA gave a better result for the tuned models. The downside to both the CPH modelling and PCA methods is that they are both relatively simple approaches. Ranking by C-index does not consider interactions between features, and PCA only produces linear combinations of genes, not allowing for complex non-linear interactions between these features. An alternative to PCA could be the use of t-distributed Stochastic Neighbour Embedding (t-SNE), which can reduce dimensionality while representing non-linear relationships in the data. However, this method is far less interpretable than PCA, as no associated loadings matrix exists, and therefore no direct link to the genes that drive the ML algorithm.

During evaluation of RSF model results, feature weights were shown and discussed. These features often had very small weights. This is not concerning if taken in isolation. However, the combination of small weights, minimal differences between the weights of each gene, and large confidence intervals meant that interpreting which genes were most important became very difficult. This is in sharp contrast to the classification modelling in Chapter 6, where *NOD2* could be determined as the most important in comparison to other genes for the disease subtype classifier. Using a method such as SHAP [453] to further elucidate the impact of each gene on modelling decisions would have been beneficial. However, the SHAP python package as it is currently constructed has limited compatibility with survival ML models, and can only provide this information for a specified time point. Given the sparse nature of the data, that is that there are few individuals present at each time-point, especially when broken down by stricturing endotype status, this analysis was inappropriate for a cohort of this size.

Throughout the tuning of the RSF models, there was a trend for the Bayes Search algorithm to choose hyperparameter values that would simplify the tree models created. This meant that maximum depth was small to create shallow trees, and minimum samples per leaf and split values were relatively large. The outer fold C-indices achieved were often better than random, and reached values of 0.6 or greater. There was better generalisation from the inner fold to the outer fold of the nested cross validation than observed for the stricturing endotype classification in chapter 6. This is potentially because the cross-validation schemes changed to boost the sample size of each fold (chapter 6 used 7-fold outer CV and 5-fold inner CV, whereas here 5-fold outer CV and 3-fold inner CV was used). Therefore, sample size was unlikely to be directly responsible for the creation of these simplified RSF models. Instead, the complexity of the data may have resulted in a need to simplify the modelling using the hyperparameters. This is supported by the reduction of overfitting (shown in the training AUC over time) in hyperparameter tuned models.

The feature selection method utilised in this chapter highlighted one of the main limitations of genetic data, even when collapsed into a GenePy matrix, the sparsity of the data. The CPH modelling required linearly independent columns, but even with this condition fulfilled the model struggled to converge on a solution with a very sparse dataset. The format of GenePy as a per-gene score is undoubtedly better than representing each variant, as this data would be even more sparse. However, the implementation of the CADD Phred cut-off (variants with a CADD Phred score less than 15 removed from the dataset) does then make the GenePy matrix sparser, while admittedly reducing dataset noise. Genomic data processing is a balancing act of ensuring that false variant calls and benign variation is not overtaking true pathogenic variation, while not having too many hard filters in place such that some of that pathogenic variation is excluded. A possible solution to the data sparsity issue would be to add GenePy scores together in a biologically sensible manner. This could be done by adding GenePy scores together if the genes associated proteins form a complex [387], or addition of GenePy scores across sections of pathways. Another possible alternative is to utilise network analysis to identify which scores should be added together. In any method which adds GenePy scores, a normalisation method needs to be used, so that GenePy scores with a small range are not overshadowed by longer genes that accrue more variation. A downside of adding many GenePy scores together is a loss in granularity with regard to the specific genes that drive patient stratification.

The issues of data sparsity, oversimplified algorithms, and modest performance could be solved by increases in dataset size. Almost all RSF models generated in this chapter showed signs of overfitting on the training dataset, suggesting that sufficient data was not present to create models that were generalisable. Research programs such as Gut reactions [382] and UK BioBank [487] are rich sources of genomic data. UK BioBank in particular is very large, and a recent data release contained the sequencing data of approximately 450,000 individuals, of which 4,614 have an IBD diagnosis (2,907 with CD). It is almost certain that with an increase in dataset size, the RSF algorithm would have more power to detect genomic signals, boosting performance. In addition, larger datasets open up opportunities to employ the previously mentioned Cox-nnet [483], and other deep learning survival analysis methods like SurvNet [488] and DeepHit [489]. With more power, it is also more probable that the hyperparameters chosen during tuning will create more complex ML models. A larger dataset has the potential to be less sparse, as there is an increased likelihood of a higher proportion of individuals in the cohort exhibiting rare variation that will meet the CADD cut-off. However, it is also likely that more *de novo* variants will be observed, which will increase sparsity. It is therefore difficult to say how much a larger dataset would mitigate GenePy matrix sparsity.

In this chapter, a method utilising binomial probabilities was developed that could be used to reduce the sparsity of the dataset. This method can be applied to any future modelling where GenePy scores are utilised. Its application can also extend beyond GenePy matrices, and be used as a filtering step for any dataset where sparsity is a limiting factor in computational modelling. This issue is likely to become ever more prevalent as ‘big data’ is more commonly leveraged. While the NOD-signalling pathway gene panel gave the highest average testing AUC and C-index from all the RSF models constructed, because these performance metrics are only modestly good, it is difficult to draw conclusions regarding an association between the genes selected and the aetiology of the stricturing endotype. Confidence in the relationships between the genes selected and the outcome is only as strong as the testing performance achieved by these models. As the genes contained within the NOD-signalling pathway gene panel are also included in the IBD gene panel, the results here do provide some corroboration with the results in Chapter 6, where the IBD gene panel was found to produce the best performing stricturing endotype classifier. The best performing gene panel here was the smallest one, and not the IBD panel, and this is potentially a reflection of the difficulties surrounding modelling when the number of features exceeds the number of genes, especially when using an ML pipeline that utilises a regression-based algorithm. It is difficult to know whether this model performed more poorly than the stricturing classifier in Chapter 6 because of the construction of the ML problem (binary classifier versus regression), or because of the increased number of genes included in the stricturing endotype classifier that utilised the IBD panel providing more information for classification. In order to make confident predictions regarding a set of genes contributing to a stricturing endotype, the RSF model must achieve better testing metrics. Having relatively few patients per follow-up timepoint (see Figure 49), has an impact on the RSF model's ability to spot patterns and be generalisable to new datasets. A limitation common to many other studies, restricted follow-up time, was not present here. Rather, there is a need for more patients at already existing timepoints. This need was highlighted with the aforementioned RSF model instability, where the data included in the training dataset was varied in the initial modelling, and changes in the make-up of the training data changed the model features, and the performance on testing data. Although there was a trend in testing AUC increasing over time in the initial modelling in Section 7.3.1, this trend was not consistently replicated. These different trends in testing AUC over time for each RSF model also indicate a lack of generalisability of RSF models to different datasets. Finally, the small feature weights in all RSF models produced here follows a trend observed for the subtype classifier in Chapter 5, and the stricturing endotype classifier in Chapter 6, that there is no subset of genes chosen through feature selection in the ML pipelines that can be used to infer any strong relationships between a gene subset and the modelling outcome. In this way, the feature selection seen throughout is reminiscent of the current biological understanding of IBD, that this is a complex, polygenic disease.

The results of this chapter, combined with the RSF results on clinical data achieved by Ungaro et al. [485], show that there is potential in creating a survival model for the stricturing endotype. Key to the future development of this modelling will be a larger cohort, and the integration of clinical and 'omic data. These types of survival methods can be more interpretable for clinicians, as the results of RSF modelling can be plotted as Kaplan-Meier curves for individual patients. This makes the RSF a good fit as a decision support tool, because the predicted risk is straightforward to interpret.

Chapter 8 Systematic review of artificial intelligence and machine learning for inflammatory bowel disease: a reassessment of the field

Chapter summary – the systematic review in this chapter follows a similar search strategy as the review in chapter 2, but has been updated (search performed May 2021) and focuses on inflammatory bowel disease. Once again, the aim was to evaluate the popular research questions for ML, which algorithms were most frequently used, and what types of data were common. In addition, comparisons are made between the popular approaches discovered in chapter 2's systematic review, and the popular methods that have emerged since. There is also a focus on how assessment and construction of machine learning pipelines has changed after approximately 30 months.

Chapter contributions – systematic search performed by Imogen Stafford. Imogen Stafford and Enrico Mossotto were first and second reviewers, respectively, for the assessment of study abstracts. Imogen Stafford and Enrico Mossotto gathered data from papers (each did full read throughs of 50% of the total papers to be reviewed). All further analysis and data synthesisation performed by Imogen Stafford. Sunburst plot generated with the assistance of Mark Gosink.

8.1 Introduction

Since the original systematic review search in Chapter 2, which focused on AI and ML applications for some of the most common autoimmune diseases [490], interest in AI for personalised medicine has only continued to grow. Among other initiatives in this field, the digital healthcare branch of the National Health Service (NHSX) announced the first round of recipients for the AI in Health and Care Award in September 2020 [491]. The investment of £140 million into this scheme highlights this continued interest. In light of this, the application of AI and ML to IBD was re-evaluated. There have been other, more recent systematic reviews in this area, namely Nguyen et al.'s review of ML for IBD diagnosis and prognosis [492], and Tontini et al.'s evaluation of artificial intelligence uses for gastrointestinal endoscopy [493]. The aim of this systematic review was to re-evaluate the common data types, applications and methods used ML for IBD, and compare these results to the original data gathered in Chapter 2. In addition, this review is intended to place the results of the previous chapters into context with other research being conducted in this field, with emphasis on the most recent research (studies published after Chapter 2's systematic review). In this review, the systematic search is broad, so that trends in ML for IBD can be assessed. There is a need to identify the strengths and weaknesses of this interdisciplinary field, and the emerging approaches that could be beneficial for IBD patients.

8.2 Methods

8.2.1 Literature search

An electronic literature search was performed using two databases available through OvidSP: MEDLINE(R) and Epub Ahead of Print, In-Process, In-Data-Review & Other Non-Indexed Citations and Daily 1946, and Embase 1974. In the previous systematic review, the Computers & Applied Sciences Complete database on EBSCO was also used to attempt to capture computational sciences studies that may be geared more towards method development, but still use clinical data. However, many of these research studies were also captured by the OvidSP databases, and they often used synthetic data. Due to this previous experience, it was judged that a search of MEDLINE(R) and Embase would be sufficient. A similar search strategy was followed to that in Chapter 2, with search terms being combined with Boolean operators. This search was completed on the 6th of May 2021, constructed as follows: ("machine learning" OR "artificial intelligence") AND ("Crohn* Disease" OR "Ulcerative Colitis" OR "Inflammatory Bowel Disease"). Any research paper with these terms contained in the title, abstract and/or subject headings would be captured in the list of records.

8.2.2 Inclusion and exclusion criteria

As with the literature search structure, the same inclusion and exclusion criteria was applied to this review as used in Chapter 2. The only exception to this is the disease-specific criteria: studies that applied ML to IBD, or an IBD subtype, were included (as opposed to the broader scope of autoimmune disease in earlier work). Studies that utilised ML for analysis of non-IBD complications on an IBD cohort were also included. Usually these types of studies centre around comorbidities that IBD patients could be more susceptible to, for example osteoporosis [494]. Studies not written in English, or that were published before 2001 were excluded. Research had to be performed using human patient data to be included. Additionally, records that were not peer reviewed, or were not original research articles were excluded, leading to no assessment of the following publication types during screening (as categorised by OvidSP): conference abstracts, conference review, editorial, erratum, journal article comment, journal article review, letter, letter comment, note and review. After these initial exclusions, two reviewers independently assessed each record's abstract to determine if it should be included or excluded. Where consensus could not be reached based on the abstract, the full text was read by both reviewers to decide on its inclusion. This systematic review conforms to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) standards [177].

8.2.3 Data collection and visualisation

The following data items were collected for each study that met the criteria: the clinical task for which ML was applied; the type of ML (supervised or unsupervised); all ML algorithms trialled by the researchers; the best performing ML algorithm; sample size; cohort type (IBD, UC or CD); data type; the best results achieved; whether a training and testing split was used; if other cross-validation was used; if the model was applied to independent test data, and the year of publication. Where the e-publication date was superseded by a print date, the e-publication date was used.

Summaries of sample sizes for each ML method, and the popularity of ML methods over time, were visualised using ggplot2 in R [178, 179]. For the sample size graphic, an ML method was counted as being used if it was recorded as a method in the research paper, even if that ML method did not generate the model with the best performance. For clarity, ML methods were sorted into groups. For example, ridge regression and logistic regression were both included in the regression group. If multiple methods from the same ML group were used in a paper, the method group was only counted once to avoid skewing the data. Some papers investigated multiple IBD research questions with different sample sizes. In these cases, each task that applied ML was counted as a separate entry. All ML method groups where sufficient data was present for a boxplot ($n \geq 5$) were included. The same ML groups were used for the visualisation of the use of different ML groups over time.

8.3 Results

Through conducting the systematic search, 409 records were identified. Of these, 135 records were identified as duplicates and subsequently removed. Through application of the study inclusion and exclusion criteria that specified article type, publication year and language, 153 records were removed. Then, the abstracts of 121 articles were screened, and 33 were subsequently excluded. A further 9 research papers were excluded after a reading the full text (Figure 69). The focus of this review was to present summary statistics for the 78 studies, detailing the most popular ML methods, applications and data types used, sample sizes, and the frequency of the implementation of cross-validation. The chosen ML models and data for each type of clinical ML task are detailed in Table 54.

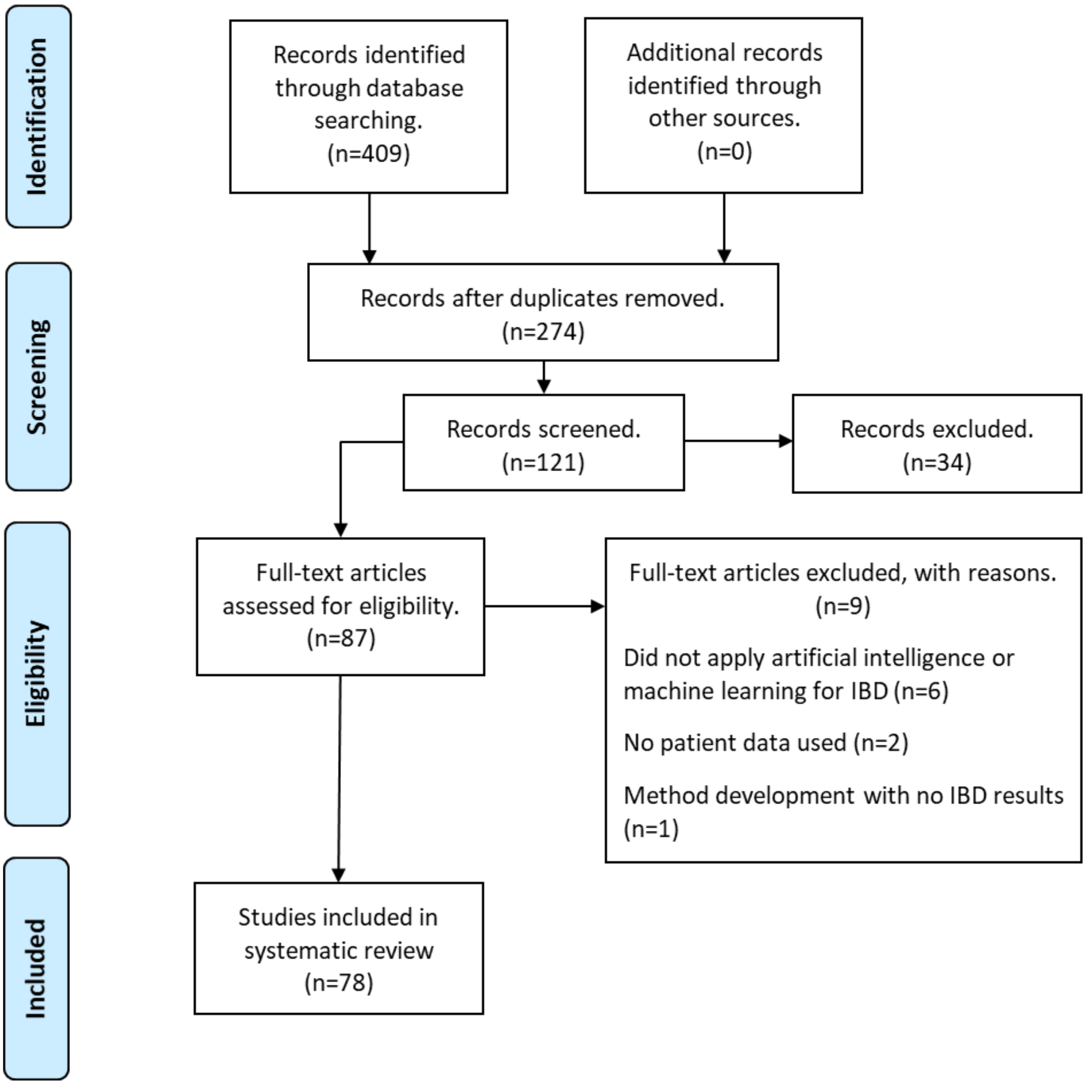

Identification
Screening
Eligibility
Included
Records identified through database searching. (n=409)
Additional records identified through other sources. (n=0)
Records after duplicates removed. (n=274)
Records screened. (n=121)
Records excluded. (n=34)
Full-text articles assessed for eligibility. (n=87)
Full-text articles excluded, with reasons. (n=9)
Did not apply artificial intelligence or machine learning for IBD (n=6)
No patient data used (n=2)
Method development with no IBD results (n=1)
Studies included in systematic review (n=78)

Figure 69 PRISMA flowchart of the number of records found, reviewed and excluded at each stage. Records were first identified through database searches, and the unique records from these searches assembled. After applying the initial exclusion criteria (record type, English only, year of publication), the remaining studies were screened by reading the abstract. Research that could be identified from the abstract as not meeting inclusion criteria was excluded. After a full text read, some reports were excluded (with reasons).

Table 54 Summary of ML models chosen as most optimal for the clinical task, and the types of data used (ML models and data types sorted alphabetically).

Task	Number of Studies	Chosen ML Models	Data Types Used
Disease Course	22	Bayes Network, Boosting, Decision Tree, Hierarchical Clustering, Neural Network, Partial Least Squares Discriminant Analysis, Random Forest, Regression, Support Vector Machine	Clinical, Gene Expression, Genetic, Imaging, Metabolomic, Metatranscriptomic, Microbiome
Diagnosis	18	Boosting, Hierarchical Clustering, Neural Network, Random Forest, Regression, Support Vector Machine	Gene Expression, Genetic, Imaging, Metabolomic, Microbiome
Disease Severity	16	Bayes Network, Boosting, Decision Tree, Hierarchical Clustering, Intelligent Monitoring, Neural Network, Regression, Support Vector Machine	Clinical, Gene Expression, Genetic, Imaging, Protein Biomarkers
Disease Subtype	8	Boosting, Hierarchical Clustering, Random Forest, Similarity Network Fusion Clustering, Support Vector Machine	Clinical, Gene Expression, Metabolomic, Microbiome
Treatment Response	7	Neural Network, Random Forest	Clinical, Gene Expression, Microbiome
Risk of Disease	6	Ensemble Model, Random Forest, Regression	Clinical, Gene Expression, Genetic
Patient Clustering	4	Gaussian Mixture Model, Hierarchical Clustering, Latent Dirichlet Allocation, Neural Network	Immunoassay, Metagenomic, Online Posts, Questionnaire
Medication Adherence	1	Support Vector Machine	Clinical
Metabolite Abundance	1	Sparse Neural Encoder-Decoder Network	Metabolomic, Microbiome
Identification of Patients	1	Natural Language Processing	Clinical

The majority of included studies employed supervised ML, with four papers using unsupervised methods [391, 495-497], and five utilising both ML types [254, 498-501]. Many studies compared multiple ML algorithms before selecting the optimal method, and some researchers investigated using ML for multiple clinical problems. Three main areas of investigation using ML methods were identified: diagnosis (23%) [262, 272, 437, 440, 441, 498, 502-513], disease course (28%) [432, 433, 435, 498, 509, 514-530] and disease severity (21%) [255, 271, 434, 485, 501, 531-541]. Diagnosis was defined for this review as classification of IBD patients (or a subtype) and controls. ML for disease course involved applications of algorithms to remission, relapse and surgery. Studies on IBD activity, or predicting the development of complications, were included in the disease severity category. The most prevalent ML method was random forest (47%); regressions, neural networks and support vector machines were also utilised often (31%, 28% and 27%, respectively, Figure 70). Other tree-based methods were trialled in 22% of research papers (13% tree-based boosting, 9% decision trees). Percentages here sum to over 100% because multiple methods were tested per study.

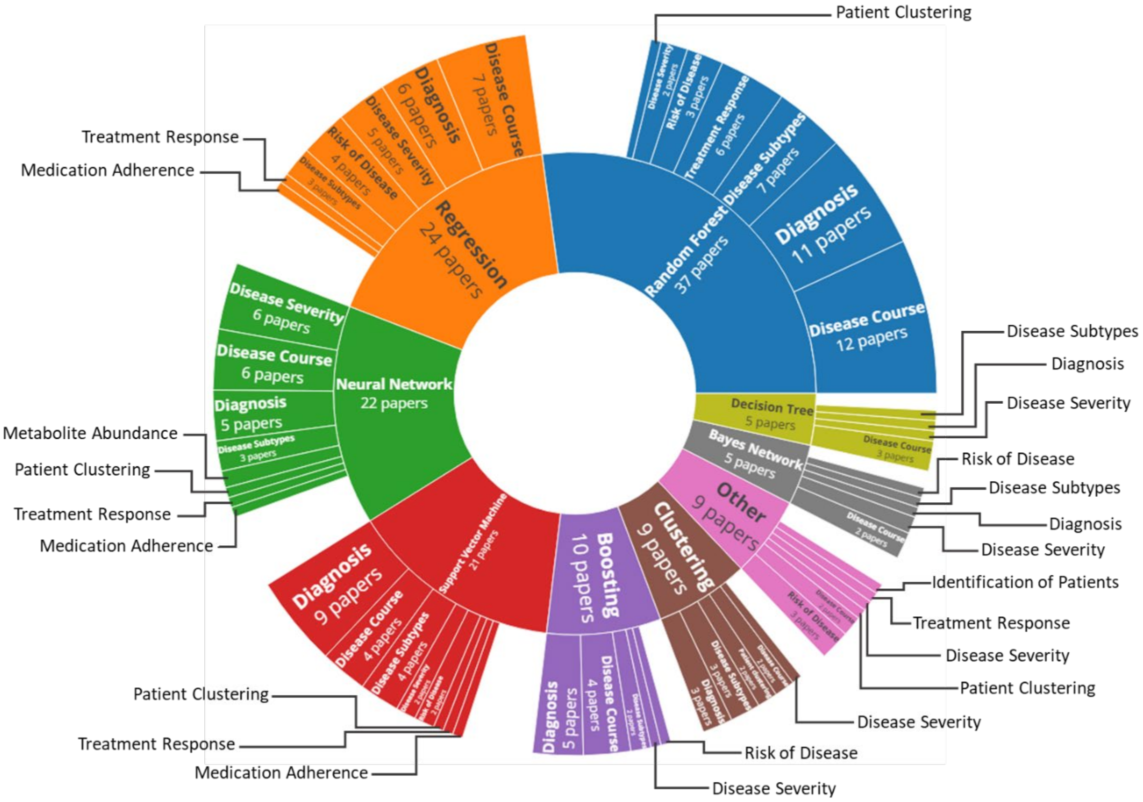


Figure 70 Sunburst of machine learning methods and the classification tasks used in conjunction with them. Categories labelled with black text have one paper associated with them.

In terms of data types, the most commonly used were clinical (41%) and microbiome (23%) datasets. The median sample size of data used for training and testing, which excludes datasets that were used for additional validation, was 263 (range 12 – 7,4000,000). Figure 71 details sample sizes for each ML method group. Only 5% of studies utilised validation datasets in addition to using the expected training and testing datasets [496, 511, 518, 542]. Separately, seven studies chose to train ML algorithms with cross-validation on one dataset, and then test the trained model on an external, independent validation dataset [440, 505, 522, 532, 534, 538, 543]. Not all studies investigated the same type of cohort: 27 had a dataset of only CD patients [271, 433, 435, 437-441, 485, 496, 500, 501, 508, 511, 514-516, 518, 521, 524, 525, 533, 536, 544-547], 15 used only UC data [255, 260, 507, 517, 519, 523, 526, 529, 532, 534, 535, 537, 540, 541, 548], and the remaining 36 used a mix of CD and UC data, either labelled as their subtype or as IBD [254, 262, 272, 391, 432, 434, 436, 495, 497-499, 502-506, 509, 510, 512, 513, 520, 522, 527, 528, 530, 531, 538, 539, 542, 543, 549-554]. Half of the research on UC only data focussed on the prediction of disease activity from endoscopy data, but for CD data the applications of ML were more varied. More summary information about each paper included in the systematic review can be found in Supplementary Table 4.

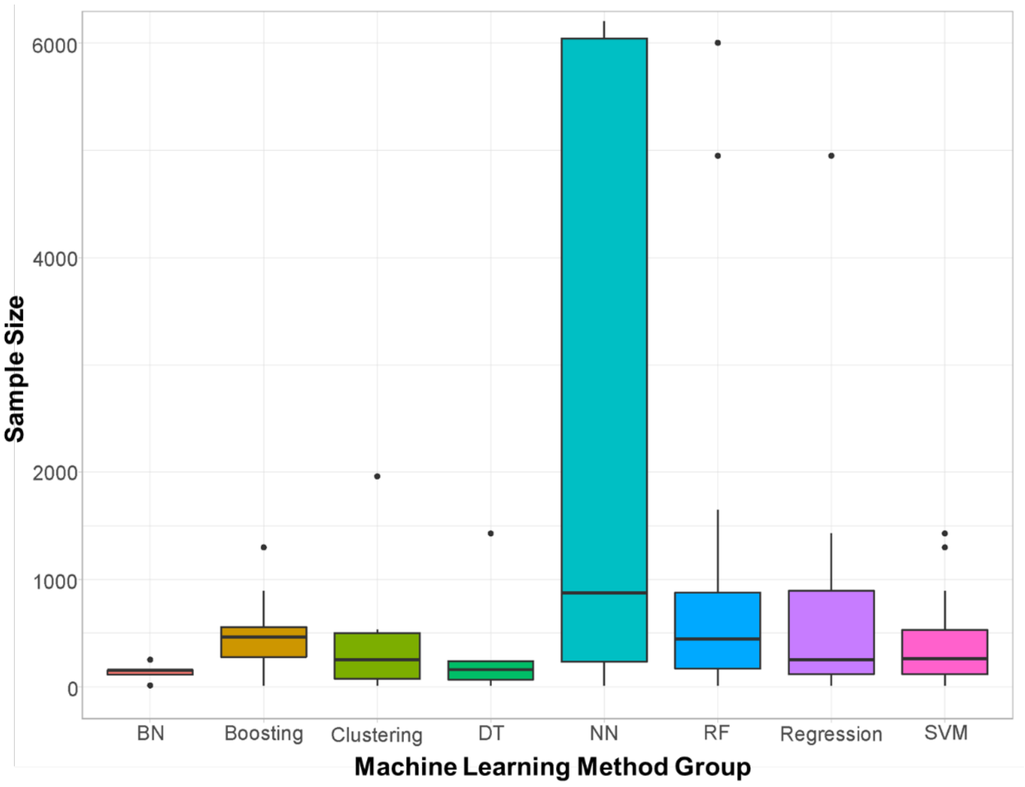


Figure 71 Sample sizes used for each group of machine learning methods. BN = Bayes Network, DT = Decision Tree, NN = Neural Network, RF = random forest, SVM = support vector machine. Note that 10 outlier entries (sample sizes 20,368-7,400,000) in the Neural Network category have been excluded from the visualisation.

8.3.1 Comparison with the systematic review of AI for autoimmune disease

The systematic search for the previous review was completed on the 18th of December 2018. Therefore, to assess the changes in the field since, studies published before and during 2018 were compared to those published from 2019 to the 6th of May 2021. In the intervening 2.3 years, 53 research studies have been published (an increase of 68%). One method's usage has rapidly increased in comparison to earlier work: neural networks, the deep learning-based method, were used in 21 studies from 2019 onwards, and only in one study prior to this. This method was often applied to imaging datasets, and as such this increase in neural network usage coincides with an increase in those types of datasets, (4% 2007-2018, 18% 2019- May 2021), particularly colonoscopy imaging data. Support vector machines, random forests and regression-based methods were all popular in both time periods (ML group method usage by year in Figure 72). The studies from 2019 onwards used two data types as variables in their ML pipelines more frequently (8% 2007-2018, 17% 2019-May 2021), almost always by using clinical data alongside another data type. Since 2018, the median sample size has not increased, as it was 273 prior to 2019, and is 257.5 for the set of studies published between 2019 and May 2021. In both time periods diagnosis was a popular application, but previously treatment response was more popular (24% 2007-2018, 1.8% 2019-May 2021), and now applications of ML to disease course questions is the most popular application (12% 2007-2018, 35.8% 2019-May 2021).

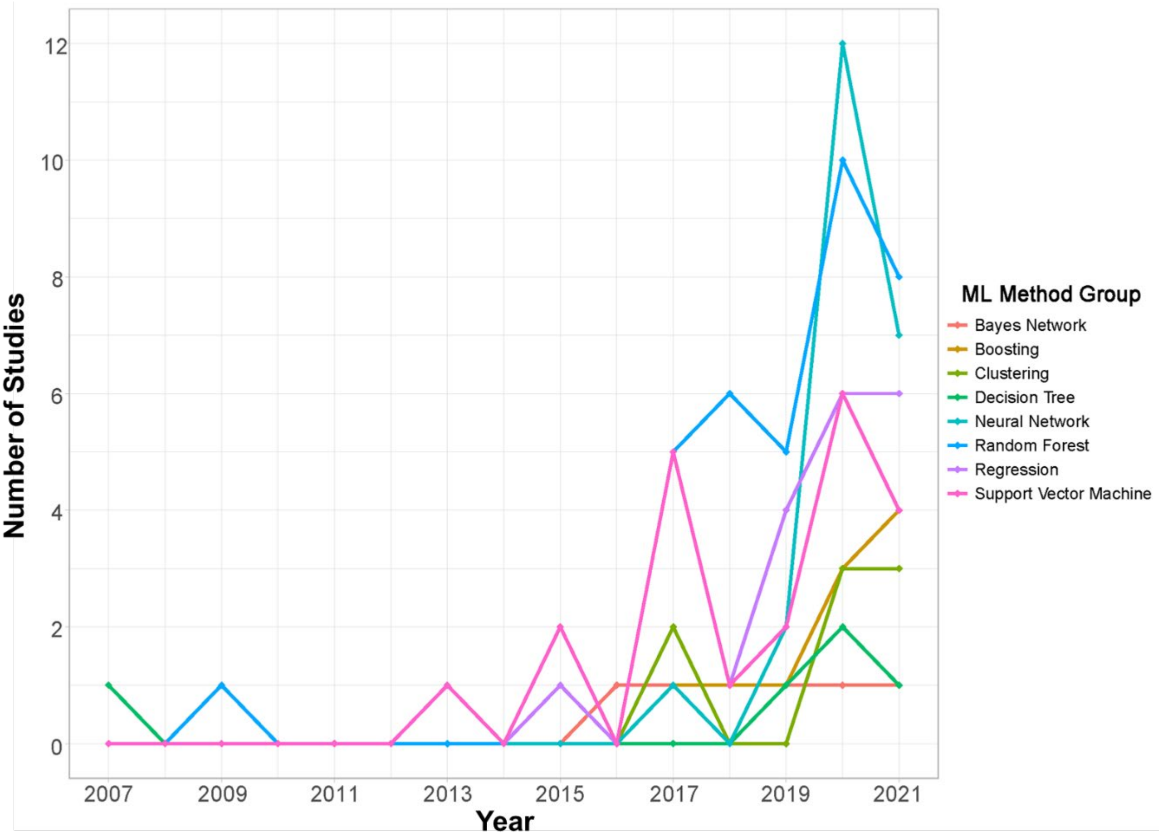


Figure 72 Implementation of machine learning methods over time, incomplete data for 2021 (includes data up to and including May 2021).

8.4 Discussion

The explosion of ML usage for IBD detailed here reflects the wider interest in AI for medicine. There was a great deal of heterogeneity in the clinical tasks ML was applied to, the data used and the ML methods themselves, but there was also heterogeneity beyond what was documented for this review. ML pipelines vary hugely. Some used feature selection and some did not, and when feature selection methods were used there were a wide array of methods. Additionally, hyperparameter tuning was not undertaken in all research papers, which can potentially affect model performance. Furthermore, datasets can be processed in different ways prior to ML. This is particularly relevant for ‘omics datasets. These factors mean that it is not possible to conclude that one approach is superior. In fact, it is probable that the superior approach will be different, from one combination of clinical task and data type to another. In future systematic reviews, it may be beneficial to document the data transformation, feature selection, and hyperparameter tuning methods used alongside the ML method. All three can affect the generalisability and utility of ML algorithms. These aspects may also become more pertinent to the translation of models to the clinical setting if in the future particular ML methods are shown to be superior for specific data types.

It is possible that the search strategy used may have led to the exclusion of some studies from the review, as only free text searches were performed, and Medical Subject Headings (MeSH) were not. However, when the subject heading “machine learning” was exploded, (OvidSP’s term for including all terms related to the search term) the only additional sub-heading was “support vector machine”. Therefore, exploding out this MeSH term could have biased the search strategy, as it would only identify additional papers using this specific method. Another possible approach to this search strategy could have been to search for each main ML method separately, alongside the broad “artificial intelligence” and “machine learning” terms. However, performing searches for “regression” would have produced too many records to assess. No risk of bias assessment was performed for each paper here, as there is no clear equivalent of PROBAST (Prediction model Risk of Bias Assessment Tool) for the evaluation of ML modelling. The construction of such a tool would be useful for the assessment of ML modelling, particularly one geared towards ML applications for clinical settings. A discussion of the reduction of bias in ML models shares many similarities with the discussion of the generalisability of ML models, as a model that is not generalisable would be biased towards the dataset it was created with. Therefore, ways to assess generalisability, and overall improvements that could be made to ML pipelines are addressed in the remainder of the discussion.

Tree-based ML methods were abundant in the studies that met the criteria for inclusion: one or more of random forests, decision trees, and tree-based boosting were implemented by 55% of research papers. Trees can be advantageous for clinical applications, as they are highly interpretable. The use of boosting and random forest (which contains many bootstrapped trees), exchanges some of this interpretability for an ML model that is less prone to overfitting, and so more generalisable. Random forests are also well known as an algorithm that can leverage non-linear relationships between dataset features. This popularity is not inherently troublesome, but a deficit of comparisons of different ML methods to random forest where appropriate, or an absence of reporting these comparisons could make developing these ML pipelines for clinical applications more challenging.

Overall a good range of informative metrics regarding the performance of ML algorithms were reported in the included studies. This was a much-needed change from the previous review, where it was noted that some studies only reported accuracy (although this observation was made for all autoimmune disease ML models, not IBD specifically). Reporting metrics such as sensitivity, specificity, AUC and F1 was particularly important where data with imbalanced classes was used. This is because high accuracy can mask poor prediction for individuals in the minority class, and often for these clinical problems the minority class will be the class with the IBD complication, or more severe disease. Some studies attempted to correct imbalanced data through algorithm weighting [549], or by oversampling the minority class [433, 537]. However, some studies did not explicitly address their imbalanced data. While datasets are often imbalanced (because this represents the nature of the patient population) it is always key to evaluate if enough samples from each class were present in the training data so that well-informed predictions could be made by the ML algorithm for both the classes. Another potential ML pipeline issue identified in some of the research studies included here was applying feature selection to the whole dataset, as opposed to only the training dataset. By doing this, information on the characteristics of the features in the test set could leak through to the training set, biasing the ML model.

In comparison to the previous systematic review, the use of independent validation datasets alongside training and testing data did not increase. However, other interesting approaches to maximise dataset usage were observed. Some studies used a leave-one-dataset-out approach, which uses cross-validation principles to utilise many smaller datasets in ML training [508, 551], while other studies compensated for relatively small datasets by training an ML algorithm with cross-validation on one dataset and testing this model on an external dataset [440, 505, 522, 532, 534, 538, 543]. The range of total dataset sizes was vast, and some of these small sample sizes may not have been sufficient for the chosen ML method. However, it can be difficult to evaluate if the sample size is adequate because this varies depending on the ML task, as well as the method, and there is no standard power calculation for ML algorithms. Algorithms such as neural networks require more data, and as shown in Figure 71, these methods did have the largest datasets. In addition, the number of features used will also affect the sample size that is necessary to generate a generalisable model. More features will generally result in a more complex model, so for the algorithm to accurately detect these patterns larger datasets are necessary. If an ML model generalises well from training to testing data (and/or independent data), this is indicative that the dataset size was sufficient. However, it is also key to evaluate how representative the dataset is of the wider patient population. An ML algorithm may achieve good results on the testing data, but upon application to external datasets may perform poorly, because the patterns the ML model previously detected are not representative.

Diagnosis was still a popular application for the period 2019-May 2021. However, the largest cluster of papers in this latest time period examined questions surrounding disease course. This is encouraging, as there is a move away from the less clinically useful task of classifying patients from healthy controls, and a move towards solving questions that will have a larger impact on patients’ quality of life. It also suggests a trend towards more comprehensive patient data collection, with deep phenotyping data allowing for the creation of these ML classifiers. It was surprising that the median dataset size had not increased in recent years, especially as increasingly large datasets are a supposed hallmark of this current era of research. It is potentially difficult to gather such data because it requires the linking of many variables, including laboratory, clinical and ‘omics data, with the phenotypic outcome it is desirable to predict. It is also challenging to link together data from different studies, as different variables can be collected by each study. A community effort may be required to accumulate the dataset sizes necessary for high-performance, generalisable ML algorithms. Large datasets are also required for further validation. Initiatives such as UK BioBank [487] and the Gut Reaction database [382] could be instrumental in progressing the creation of new ML pipelines. By utilising these kinds of data, along with robust pipelines and generalisable ML models, this advances personalised medicine for IBD patients.

Chapter 9 Summary and future research

This thesis has sought to find and utilise novel approaches for the stratification of IBD patients. Oxidative stress and antioxidant potential assay data were investigated for correlations with patient characteristics and clinical data, and their link to variation in associated genes was also analysed. Both classification- and regression-based ML approaches were implemented to stratify patients using their whole exome sequencing data, transformed into the pathogenicity burden score GenePy. Random forest classified IBD patients by their disease subtype, and CD patients by the presence or absence of a stricturing endotype. Random survival forest was subsequently also utilised for CD patient stratification by stricturing endotype development.

During ML modelling of IBD disease subtypes, and subsequent analysis, *NOD2* was confirmed as the strongest discriminatory gene. *NOD2*'s effect on classification was amplified after the hyperparameters of the subtype model had been tuned. No other gene had the same discriminatory power in the subtype ML model, or in any other model. This disease subtype ML model achieved the best performance with an autoimmune gene panel. This included a better performance than an IBD panel containing genes known to cause monogenic forms of IBD and loci discovered through GWAS, which indicates there are potentially new susceptibility genes to uncover. A potential hypothesis for the reason *NOD2* is a strong predictor is that this is a gene known to contribute to CD development in both a monogenic [5] and polygenic [2, 3] way. This suggests that all *NOD2* variation should be considered for potential clinical investigation, and not just the most common and well known polymorphisms. Through the subtype ML modelling and subsequent explainability from SHAP values an interesting relationship between pathogenic variation and classification was revealed. Despite potentially pathogenic variation conveyed through high GenePy scores being associated with discriminating CD cases through SHAP values, the subtype model was more sensitive to identifying UC cases. This suggested that lack of genomic variation was one of the more useful signals in discriminating between the two subtypes. This also provides evidence for the argument that CD is more genetically heterogeneous than first thought, with the existence of potential genetic subgroups within this disease subtype. The CD subtype could benefit from further stratification. In relation to the stricturing endotype classifier, it was thought this would achieve a better performance. It was hypothesised that the stricturing endotype could be a genetic subgroup within the CD subtype. In fact, stricturing endotype classification turned out to be very challenging, and the stricturing and not-stricturing endotypes were indistinct because of shorter follow-up in a proportion of the cohort. It was hoped that further modelling with random survival forest could improve results, as follow-up time became integral to the model. However, overfitting and RSF model instability indicated that this modelling was limited by the size of the patient data set. Overall, AUCs achieved in the disease subtype and stricturing endotype binary classifiers were good (both achieving an AUC of 0.66 after hyperparameter tuning in Chapter 6). However, because the testing AUCs produced weren’t high enough to have high confidence in generalisability of the models, there was an inability to draw any conclusions regarding new genes and pathways that were involved in stricturing endotype or disease subtype classification. In order to investigate this in the future, unsupervised learning could be employed for gene and pathway investigations, and because this would be agnostic to clinical labels, new genetic subgroups could be found in tandem.

As exemplified in the two systematic reviews included in this thesis, this interdisciplinary field, which applies computational methods to big data for patient benefit, is rapidly evolving and dynamic. Over the relatively short period of my PhD studentship, the implementation of these algorithms has shifted from having straightforward aims of classifying healthy controls and individuals with disease, to a focus on disease course and severity. The latter models are more likely to provide useful decision support when translated to a clinical setting. In addition, there has been a shift towards the use of the deep learning method neural networks for imaging data in recent years; many of these ML pipelines were focused on disease severity prediction from colonoscopy images in UC patients. There may be an uplift in the performance of ML algorithms as datasets grow. In addition, implementation of more ensemble ML algorithms have the potential to increase performance [555, 556]. Figure 73 shows AUCs achieved by comparable ML models (testing AUC information was available) for subtype classification identified through the systematic reviews conducted, alongside the results obtained through this research.

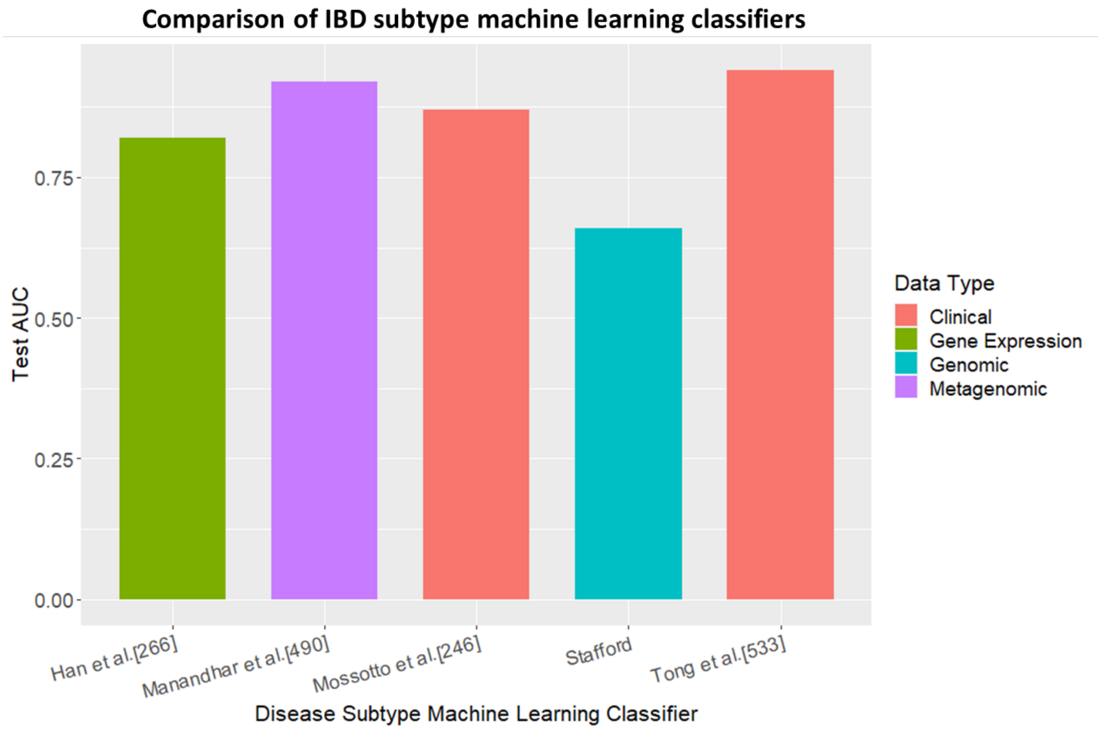


Figure 73 Bar plot of IBD subtype machine learning classifiers identified during Chapter 8's systematic review, that could be compared to the subtype classifier constructed in the previous chapters.

Alongside the strides made in methods used, and the more complex clinical research questions, there has been an increase in the availability of large genomic datasets. Genomics England data has recently become publicly available, and the number of exomes included in UK Biobank datasets has been increasingly regularly. The National Institute of Health Research Gut Reaction database is an IBD-specific resource that is now available, and provides curated data resources, including clinical and genomic data. Locally, on a smaller scale than national-level genomic cohorts, the Southampton IBD cohort has more than doubled with the inclusion of over 500 adult IBD exomes. This is, in part, due to the decreasing costs of sequencing. In tandem, the bioinformatic tools that are available to annotate these data are refined and improved, and novel software is introduced into the lexicon. This includes the 2018 update of gnomAD v.2.1, primarily used for allele frequency, followed by the gnomAD v.3 release in 2019 which was able to provide better allele frequency annotation for whole genome sequencing data than the previous release. In addition, CADD has been upgraded three times between 2018 and 2022, with the last update incorporating splicing annotations from SpliceAI [363]. There are also new ensemble pathogenicity predictors, such as BayesDel [557] and REVEL [558], which may provide more precise predictions than CADD [559]. The ability to annotate these large genomic datasets is critical to their interpretation, and any subsequent machine learning analysis. The possibilities for the analyses of these datasets, when they are growing constantly, and annotations are becoming ever more sophisticated, are incredibly exciting. However, there are also some challenges associated with these changes, and improvements that are still outstanding.

As the number of individuals included in each dataset increases, alignment, joint calling and annotation of genomic data becomes increasingly computationally expensive and intensive. As a result, new pipelines have been developed, such as Illumina and the Broad Institute's DRAGEN-GATK, which can be run utilising cloud computing. The commercial version of DRAGEN-GATK is purported to be a fast joint-caller for samples [560, 561]. One of the main advantages of cloud computing is its scalability: the amount of computational resources available can be increased to meet the demand of each processing step. Then, once processing has been completed, resources are no longer required. Therefore, many processes can be performed in parallel. Cloud computing is now necessary for some processing of big genomic data [562]. Currently, the processing and storage of genomic data is more expensive than the initial sequencing cost. With recent releases of UK Biobank data (the final release in July 2022 containing 470,000 individuals, of which over 4,600 are diagnosed with IBD), there has been a shift from being able to download this data, to it now being stored in its own research environment (UK Biobank research analysis platform) that researchers can apply to access [563]. This will save research groups the storage cost of maintaining local copies of datasets. However, this also comes with the challenge of maintaining a research environment such that tools for processing and analysing genomic data are maintained, and new bioinformatic tools can be installed and tested in the research environment.

In order to make full use of different genomic data sets as training and validation data sets for ML algorithms, it will be important to understand how technical differences in genomic data collection impacts the variants called. In order to combine two or more datasets as one training dataset for an algorithm, it is necessary to understand how differences in factors such as sequencing read lengths and capture kits will impact downstream data. This may result in patterns being detected by ML algorithms that are the result of technical, rather than biological differences. In addition, if two datasets, used for training and validation respectively, contain technical differences, it will be important to assess if an ML model's generalisability across datasets is impacted by these differences.

Machine learning results, particularly those in Chapter 7 using random survival forest, showed that data sparsity can be an obstacle to precise prediction when utilising genomic data. This exemplifies the necessity of summary scoring systems such as GenePy, which reduce the dimensionality of datasets. Sparsity will also potentially increase as the number of individuals included increases, with more rare and *de novo* variation called in data processing. The GenePy algorithm has a key advantage over other previously discussed scoring systems, because it synthesises predictors of pathogenicity together with zygosity and allele frequency. In contrast, these other methods used only zygosity, variant consequence, and gene associations present in literature [440, 441]. However, one of the main disadvantages of GenePy is that it is currently formulated to only score bi-allelic variants. As cohort sizes increase, it is inevitable that variant sites will increasingly become multiallelic. Therefore, developing GenePy into an algorithm that can be utilised for tri-allelic variation and beyond will be paramount. Another limitation of GenePy comes from the raw data, as opposed to the construction of its algorithm. With short-read, high-throughput sequencing data, the phase of called variants cannot be determined. Therefore, compound heterozygosity cannot be factored into a summary score. There have been great strides in the accuracy of long range sequencing. The size of each read – Oxford Nanopore technology averages between 10 and 30 kilo base pairs per read [564] – means that determining phase, that is whether two variants are on the same, or different copies of a gene, is possible. Phase information is not only helpful for deducing compound heterozygosity, but also for assessing how impactful multiple gene variants could be on downstream protein function. With phasing information available, it will be possible to more accurately reflect each gene’s mutational burden in scoring systems such as GenePy. Where variants that are predicted to be pathogenic occur on both chromosomes these scores could be upweighted, and the score reduced where these variants occur on only one chromosome. The quality of data used in ML pipelines understandably has an effect on the performance and generalisability of the resulting ML models. By generating GenePy scores that more accurately reflect each patient’s genomic profile, there is potential to better stratify patients with ML algorithms. Looking forward, it will be important to consider different pathogenicity predictors to CADD, as described above. In addition, new pathogenicity burden score algorithms may be developed. These could either replace GenePy entirely, or GenePy could be re-developed by incorporating methods that other scoring algorithms utilise.

The work of this thesis primarily focused on the clinical classification tasks of stratifying patients by disease subtype and by stricturing endotype. There are a number of useful clinical classifications that were not addressed. One of these could be the prediction of the penetrating endotype, where a fistula develops, leading to abnormal connections between passages of the gastrointestinal tract, or the gastrointestinal tract and the skin. As with the stricturing endotype, this can also require surgery. The main challenge with this type of approach, as discussed in Chapter 5, is mining the free text of clinical reports and letters to accurately capture each patient’s stricturing or penetrating status to train ML algorithms. A potential alternative is to utilise the standard surgery codes, Operating Procedure Codes Supplement 4 (OPCS-4) classification of interventions and procedures, that are implemented by the National Health Service. Surgery could then be used as an indicator of severe disease. Alternatively, particular surgeries can be a reasonable proxy for an endotype. For example, many patients with a stricturing endotype will have undergone a right hemicolectomy. Here, precision of phenotype is traded-off for standardised outcome data, which makes validating an ML algorithm on external data easier, and more suitable to widespread clinical implementation.

It should also be acknowledged that utilising a classification framework may not be the best approach for fully understanding the aetiology of each patient's inflammatory bowel disease. Work in Chapter 7 using random survival forest revealed that stratification of patients based on the more specific stricturing endotype was challenging for a number of reasons. Firstly, development of this endotype at any point in a patient's disease course necessitated a regression-style approach which results in more complex model construction than a binary classifier. Secondly, there was a smaller dataset, and finally the underlying genomic heterogeneity within the cohort appeared to be greater than anticipated. Therefore, unsupervised ML algorithms may be an approach whereby patients can be grouped according to their shared causal molecular mechanisms. This could enable the discovery of IBD patient subgroups that are more accurate to disease course than the traditional subgroups of CD and UC. This discovery-based approach is potentially more useful for the ultimate aim of personalised medicine, where patients can be given treatment that is tailored to their disease mechanism. For example, the small molecule JAK-inhibitor drugs, which act to inhibit the unnecessary activation of the pro-inflammatory JAK-STAT signalling pathway [457]. However, it is less suited to the interim aim of more accurate patient care by stratifying patients into groups based on their IBD phenotype, surgery likelihood, or non-response to specific treatment. Classification based methods are more suited to this stratification, and can provide clinicians decision support. Therefore, both types of ML have a role to play in understanding IBD, and stratifying patients.

Key to the future implementation of ML for patients to enable personalised medicine, will be high quality patient data. This involves: 1) initial data gathering; 2) updating data; particularly for longitudinal analyses; and 3) linking datasets together. The latter is crucial for the incorporation of any 'omic dataset into an ML pipeline for patient stratification. However, data that can be linked needs to be secured in federated systems that protect patients, while also enabling researchers to make full use of data. A federated network, where multiple datasets can be queried and analysed, without that data leaving the node it is stored in, has already been discussed and implemented for health data [565, 566]. Alongside this, education for the public needs to be delivered so that patients are reassured about the security of data storage. Education and communication with the public and clinicians alike will also be necessary for any widespread implementation of ML algorithms as decision support tools. These ML tools will also need to work for multi-ethnic societies, and this is a particular problem with genomic datasets, where the majority of recruited individuals are of European ancestry.

Machine learning modelling conducted throughout has established that there are persisting gaps in knowledge of the genomics of IBD, and a reliance on clinical phenotypes may impede the ability to uncover more detail regarding IBD genomics. The RF modelling results here strongly imply that there is considerable genomic variability, not just between CD and UC, and stricturing and not-stricturing endotype, but also intra-group genomic variability. This provides an argument for genomic investigation as standard, as this could lead to the identification of new genes and variants connected to specific IBD manifestations, particularly for all paediatric patients who are more likely to have an unusual IBD presentation [567]. While some progress towards patient stratification h

as been made here with proof of concept for genomics classifying patients by disease subtype and stricturing endotype, more needs to be done to achieve stratification and relate this stratification to clinical phenotypes. Given the genomic intra-group variability implied by the analysis here, a better approach to stratification may be to focus on individual phenotypic characteristics, for example colonic-only inflammation, or presenting with extraintestinal manifestations, and associating these with genomic subgroups uncovered through genomic investigations, unsupervised machine learning or other computational methods. As this interdisciplinary field continues to develop, it will undoubtedly be the case that novel ML algorithms will develop, and there may be a shift towards deep learning methods. However, these algorithms will not be able to reach their full predictive potential without quality input data, and deep, longitudinal patient phenotype data. This means that alongside algorithm development, our interpretation and understanding of genomics – and other ‘omic data – must improve, in order that this data can be integrated and transformed into formats where patterns can be more easily deduced by algorithms. This will enable stratified medicine, and later personalised medicine, for IBD patients, and other individuals with complex disease, using big data.

Supplementary Material

Supplementary Table 1 List of genes currently implicated in monogenic IBD, and their corresponding phenotype.

Gene	**Phenotype**
ADA [57]	Severe combined immunodeficiency
ADA2 [114]	Cutaneous findings, neurological involvement, gastrointestinal involvement.
ADAM17 [57]	ADAM17 deficiency
AICDA [57]	Hyper IgM syndrome
ALP1 [57]	IBD
ANKZF1 [57]	Colitis
ANO1 [92]	Infantile enterocolitis and monogenic IBD
ARPC1B [57]	Wiskott Aldrich syndrome-like with intestinal inflammation
BACH2 [115]	Intestinal inflammation (lymphocyte maturation defects causing immunoglobulin deficiency)
BTK [57]	Agammaglobulinemia
CARD8 [116]	Crohn’s disease
CARD9	Familial candidiasis, IBD phenotype
CARMIL2 [568]	IBD-like primary immunodeficiency
CASP8 [569]	IBD with perianal disease, stricturing and fistulising proctocolitis, deep ulcerations (T cell dysregulation, reduced B-class switched cells)
CD3γ [57]	Severe combined immunodeficiency
CD40LG [57]	Hyper IgM syndrome
CD55 [119]	Primary intestinal lymphangiectasia
COL7A1 [57]	Dystrophic epidermolysis bullosa
CTLA4 [120]	Crohn’s disease

CYBA [57]	Chronic granulomatous disease
CYBB [57]	Chronic granulomatous disease
CYBC1 [121]	Chronic granulomatous disease manifesting as colitis (reduced expression of NADPH oxidase subunit *NOX2*)
DCLRE1C [57]	Severe combined immunodeficiency, Omen syndrome (Crohn’s disease-like inflammation)
DKC1 [57]	Hoyeraal-Hreidarsson syndrome, colitis
DOCK2 [122]	Combined immunodeficiency
DOCK8 [57]	Combined immunodeficiency, Hyper IgE syndrome
EPCAM [57]	Tufting enteropathy
FERMT1 [57]	Kindler syndrome
FOXP3 [57]	Immunodysregulation polyendocrinopathy enteropathy X-linked syndrome
G6PC3 [57]	Congenital neutropenia (Crohn’s disease-like inflammation)
GUCY2C [57]	T cell lymphopenia
HPS1 [57]	Hermansky Pudlak Syndrome (Crohn’s disease-like inflammation)
HPS4 [57]	Hermansky Pudlak Syndrome (Crohn’s disease-like inflammation)
HPS6 [57]	Hermansky Pudlak Syndrome (Crohn’s disease-like inflammation)
HSPA1L [57]	IBD (features of Crohn’s disease and ulcerative colitis)
ICOS [57]	IBD
IKBKG [57]	X linked ectodermal dysplasia and immunodeficiency
IL10 [57]	IBD
IL10RA [57]	IBD
IL10RB [57]	IBD
IL21 [57]	Common variable immune deficiency, IBD
IL2RA [57]	Combined immunodeficiency, Immunodysregulation polyendocrinopathy enteropathy X-linked-like syndrome

IL2RG [57]	X-linked severe combined immunodeficiency, atypical severe combined immunodeficiency
ITCH [57]	Autoimmune disease, multisystem, with facial dysmorphism (ADMFD)
ITGB2 [57]	IBD
IRF2BP2 [123]	Common variable immune deficiency
LACC1 [124]	Crohn’s disease
LIG4 [57]	Severe combined immunodeficiency
LRBA [57]	Combined immunodeficiency and autoimmunity
MALT1 [125]	Intestinal inflammation, persistent cytomegalovirus
MASP2 [57]	IBD
MEFV [57]	Mediterranean Fever, IBD
MVK [57]	IBD
NCF1 [57]	Chronic granulomatous disease
NCF2 [57]	Chronic granulomatous disease
NCF4 [57]	Chronic granulomatous disease
NFAT5 [126]	Autoimmune enterocolopathy
NLRC4 [127]	Enterocolitis with periodic autoinflammation
NOD2 [5]	CD
NPC1 [128]	Niemann-Pick disease type C1 (Crohn’s disease-like)
ORAI1 [92]	Primary immunodeficiency
OTULIN [92]	Infantile enterocolitis, monogenic IBD, primary immunodeficiency
PIK3CD [57]	P13K delta syndrome
PIK3R1 [57]	Agammaglobulinemia, IBD
PI4KA [139]	Neurological disease, with IBD, multiple intestinal atresia and combined immunodeficiency.

PLA2G4A [129]	Cryptogenic multifocal ulcerating stenosing enteritis
PLCG2 [57]	Phospholipase C-y2 defects
POLA1 [130]	X-linked reticulate pigmentary disorder
PTEN [57]	PTEN syndrome
RAG1 [131]	Severe combined immunodeficiency
RAG2 [57]	Severe combined immunodeficiency, Omenn syndrome
RIPK1 [138]	immunodeficiency and chronic enteropathy
RIPK2 [92]	Infantile enterocolitis and monogenic IBD
RTEL1 [57]	Hoyeraal-Hreidarsson syndrome
SAMD9 [140]	Immunodeficiency with prominent gastrointestinal tract involvement
SIRT1 [132]	IBD
SH2D1A [57]	X linked lymphoproliferative syndrome 1
SKIV2L [57]	Trichohepatoenteric syndrome
SLC26A3 [133]	Congenital chloride diarrhoea, epithelial barrier dysfunction
SLC37A4 [57]	IBD
SLC9A3 [57]	Congenital sodium diarrhea
SLCO2A1 [57]	Primary hypertrophic osteoarthropathy
STAT1 [57]	Combined immunodeficiency
STAT3 [134]	Early onset autoimmune disease
STIM1 [141]	Immunodeficiency 10
STXBP2 [57]	Familial hemophagocytic lymphohistiocytosis type 5
STXBP3 [135]	IBD, immunodeficiency, severe bilateral sensorineural hearing loss
SYK [142]	Chronic Colitis
TGFB1 [136]	IBD and central nervous system disease
TGFBR1 [57]	Loeys Dietz syndrome
TGFBR2 [57]	Loeys Dietz syndrome

TNFAIP3 [57]	Behcet like disorder
TRIM22 [57]	IBD (Granulomatous colitis)
TRNT1 [57]	Colitis
TTC37 [57]	Trichohepatoenteric syndrome
TTC7A [57]	Familial diarrhoea
TYMP [137]	Mitochondrial neurogastrointestinal encephalopathy (Crohn's disease-like)
WAS [57]	Wiskott Aldrich syndrome -like phenotype with intestinal inflammation
WIPF1 [92]	Primary immunodeficiency
XIAP [57]	X linked lymphoproliferative syndrome
ZAP70 [57]	Combined Immunodeficiency, severe combined immunodeficiency
ZBTB24 [57]	IBD

Supplementary Table 2 Detailed information for each study included in the systematic review of artificial intelligence and machine learning applied to autoimmune disease. Studies grouped by autoimmune disease.

AA=Alopecia Areata, ACPA = Anti-Citrullinated Peptide Antibodies, AI = Renal Pathology Acute Index, AID = Autoimmune Disease, AUC = Area under the ROC Curve, axSpA = Axial Spondyloarthritis, CeD = Coeliac Disease, CFS = Chronic Fatigue Syndrome, CGM = Continuous Glucose Monitoring, CI = Renal Pathology Chronic Index, CIS = Clinically Isolated Syndrome, COPD = Chronic Obstructive Pulmonary Disease, CD = Crohn's Disease, D-IBS = Diarrhoea-Predominant Irritable Bowel Syndrome, EDSS = Expanded Disability Status Scale, EHR = Electronic Health Record, EMR = Electronic Medical Record, FP = False Positive, GWAS = Genome Wide Association Study, HC = Healthy Controls, IBD = Inflammatory Bowel Disease, LASSO = Least Absolute Shrinkage and Selection Operator, LDA = Linear Discriminant Analysis, LH-PCR = Length Heterogeneity Profile or Fingerprint, ME = Myalgic Encephalomyelitis, MF = Mycosis Fungoides, MFI = Motor Function Impaired, MFP = Motor Function Preserved, MLP = Multilayer Perceptron, MRI = Magnetic Resonance Imaging, MS = Multiple Sclerosis, OA = Osteoarthritis, OND = Other neurological diseases, P = Psoriasis, PAFS = Psoriasis and Psoriatic Arthritis Follow-up Study, PAPS = Primary Antiphospholipid Syndrome, PPMS = Primary Progressive Multiple Sclerosis, PRMS = Progressive Relapsing Multiple Sclerosis, PsA = Psoriatic Arthritis, PsC = Cutaneous-only Psoriasis, PSC = Primary Sclerosing Cholangitis, PsV = Psoriasis Vulgaris, RA = Rheumatoid Arthritis, RBC = Red Blood Cell, RF = Random Forest, RSME = Root Mean Square Error, RRMS = Relapsing Remitting Multiple Sclerosis, SLE = Systemic Lupus Erythematosus, SNP = Single Nucleotide Polymorphism, SpA = Spondyloarthropathy, SPMS = Secondary Progressive Multiple Sclerosis, SSc = Systemic Sclerosis, SVM = Support Vector Machine, T1D = Type 1 Diabetes, T2D = Type 2 Diabetes, UC = Ulcerative Colitis, VOC = Volatile Organic Compound.

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Multiple Sclerosis								

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Briggs et al. 2019 [180]	No	Disease Progression	Supervised	Multivariable Regression	N=1515	Clinical, Survey and Genetic Data	.	10-fold cross-validation
Ahmadi et al. 2019 [181]	No	Diagnosis	Supervised	Neural Network	N=12 (n(MS)=5, n(HC)=7)	Clinical Data	Colour task: Accuracy=91%, Sensitivity=83%, Specificity=96%. Direction Task: Accuracy=90%, Sensitivity=82%, Specificity=96%.	Leave-one-out cross-validation
Zhang et al. 2019 [182]	No	Disease Progression	Supervised	Random Forest	N=84	MRI Data	Shape Based: AUC=0·85, Sensitivity=0·94, Specificity=0·5. Shape based with lesion segmentation tool: AUC=0·82, Sensitivity=0·95, Specificity=0·33	3-fold cross-validation
Zurita et al. 2018 [183]	No	Diagnosis	Supervised	Support Vector Machine	N=150 (n(RRMS)=104, n(HC)=46)	MRI Data	RRMS vs HC: Accuracy=87·8%, Precision=89·7%, Sensitivity=88%, Specificity=87·6%. RRMS (EDSS > 1·5) vs HC: Accuracy=88·6%, Precision=91·6%, Sensitivity=87·5%, Specificity=89·8%.	10-fold cross-validation
Wang et al. 2018 [184]	No	Diagnosis	Supervised	Neural Network	N=1357 (n(MS)=676, n(HC)=681) images. N=64 (n(MS)=38, n(HC)=26) patients	MRI Data	Accuracy=98·77, Precision=98·75, Sensitivity=98·77%, Specificity=98·76%	Hold-out validation
Neeb et al. 2018 [185]	No	Diagnosis	Supervised	k Nearest Neighbours	N=97 (n(MS)=52, n(HC)=45)	MRI Data	Data not affected by motion: False prediction rate=16·3%. All data: False prediction rate=25·5%	Leave-one-out cross-validation
Lotsch et al. 2018 [186]	No	Diagnosis	Supervised and Unsupervised	Emergent self-organising maps, Random Forest	N=403 (n(MS)=102, n(HC)=301)	Lipid Marker Data	ESOM balanced accuracy=98%. Random forest: AUC=100%, Area under the precision recall curve=98·87%, Balanced accuracy=100%, Sensitivity=100%, Specificity=100%	Nested cross-validation
Tacchella et al. 2017 [187]	No	Disease Progression	Supervised	Random Forest/Human Rating Hybrid	N=84	Clinical Data	AUC=0·725 (180 days), 0·694 (360 days), 0·696 (720 days)	Leave-one-out cross-validation

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Lopez et al. 2018 [188]	No	Disease Subtype	Unsupervised	Agglomerative hierarchical clustering algorithm	N=191	SNP Data	Rand Index=0·96	10-fold cross-validation
Supratak et al. 2018 [189]	No	Risk of Disease	Supervised	Support Vector Regression	N=32	Gait Speed Data	R-value=0·98	. (Individual models)
Sacca et al. 2018 [190]	No	Early Diagnosis	Supervised	Random Forest or Support Vector Machine	N=37 (n(RRMS)=18, n(HC)=19)	MRI Data	Accuracy=85·7%, Sensitivity=100%, Specificity=66·7% (SVM and RF)	5-fold cross-validation
Mowry et al. 2018 [191]	No	Risk of Disease	Supervised	Logistic Regression	N=6552 (n(MS)=3276, n(HC)=3276)	Clinical/Survey and Genetic (HLA) Data	.	10-fold cross-validation (tuning parameter only)
Yoo et al. 2018 [192]	No	Early Diagnosis	Supervised and Unsupervised	Deep Learning, LASSO and Random Forest	N=99 (n(RRMS)=55, n(HC)=44)	MRI Data	AUC=88·0% Accuracy=87·9% Sensitivity=87·3%, Specificity=88·6%	11-fold cross-validation
Kiiski et al. 2018 [193]	No	Disease Progression	Supervised	Machine Learning approach with Penalised Linear Regression	N=78 (n(MS)=35 (22 RRMS, 13 SPMS), n(HC)=43)	Clinical Data	Cognitive functioning: r-value 0·35 (baseline), 0·44 (13 months). Processing Speed and Working Memory: r-value 0·27 (baseline), 0·39 (13 months)	10-fold cross validation, nested cross validation
Fiorini et al. 2015 [194]	No	Disease Subtype	Supervised	Ordinary Least Squares Regression or Regularised Least Squares Regression	N=457 (n(RRMS)=170, n(SPMS)=205, n(PPMS)=68, n(PRMS)=8, n(Benign)=6)	Clinical Scales, Patient Reported Outcomes (anthropometric and questionnaires) Data.	Accuracy=78·32 (Ordinary least squares), 78·24 (regularised least squares), F1 score=0·701 (Ordinary least squares), 0·702 (regularised least squares)	Hold-out validation, testing set
Zhong et al. 2017 [195]	No	Disease Progression	Supervised	Support Vector Machine	N=72 (n(MFP)=26, n(MFI)=25, n(HC)=21)	MRI Data	HC vs MFI: AUC=0·9448, Accuracy=88·34%, Sensitivity=96·00%, Specificity=85·71%. HC vs MFP: AUC=0·8416, Accuracy=84·16%, Sensitivity=88·46%, Specificity=85·71%.	Leave-one-out cross-validation

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
							MFP vs MFI: AUC=0·8338, Accuracy=85·61%, Sensitivity=92%, Specificity=84·62%.	
Lotsch et al. 2017 [196]	No	Diagnosis	Unsupervised	Emergent self-organising feature maps	N=403 (n(MS)=102, n(HC)=301)	Clinical (Lipid Serum) Data	Balanced Accuracy=94·6%, Sensitivity=89·2%, Specificity=100%	.
Karaca et al. 2017 [197]	No	Disease Subtype	Supervised	Convex Infinite Kernel Approach (CIKA)	N=139 (n(MS)=120, n(HC)19)	MRI and EDSS Data	Accuracy=0·8889	10-fold cross-validation
Ostmeyer et al. 2017 [198]	No	Diagnosis	Supervised	Logistic Regression Model	N=125 (n(train)=71 RRMS + 12 OND; n(val)=60 RRMS + 42 OND)	Clinical (Immune Repertoire) Data	Cross-validation: Accuracy=87% Independent Test Data: AUC=0·75, Accuracy=72%	Leave-one-out cross-validation, independent test data
McGinnis et al. 2017 [199]	No	Disease Progression	Supervised	Support Vector Regression	N=47	Gait Measurement Data	RMSE 0·14m/s	Leave-one-subject-out cross-validation
Zhao et al. 2017 [200]	No	Disease Progression	Supervised	Support Vector Machine	N=1693	Clinical and MRI Data	G0: Accuracy=0·67, Sensitivity=0·81, Specificity=0·59. G1: Accuracy=0·68, Sensitivity=0·82, Specificity=0·58. G2: Accuracy=0·65, Sensitivity=0·80, Specificity=0·57. G3: Accuracy=0·54, Sensitivity=0·52, Specificity=0·55.	10-fold cross-validation
Ion-Margineanu et al. 2017 [201]	No	Disease Subtype	Supervised	Linear Discriminant Analysis, Random Forest or Support Vector Machine	N=105 (n(MS)=87, n(HC)=18)	Clinical and MRI Data	CIS vs RR: Balanced accuracy=85%, Sensitivity=87%, Specificity=83% (SVM). CIS vs RR+SP: Balanced accuracy=92%, Sensitivity=93%, Specificity=90% (SVM). RR vs PP: Balanced accuracy=81% (SVM and LDA), Sensitivity=76%, Specificity=86% (SVM), Sensitivity=84%, Specificity=78% (LDA). RR vs SP: Balanced accuracy=87%, Sensitivity=85%, Specificity=88% (SVM)	Leave-one-patient-out cross-validation

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Kocevar et al. 2016 [202]	No	Disease Subtype	Supervised	Support Vector Machine	N=90 (n(MS)=64, n(HC)=26)	MRI Data	HC vs CIS: F-Measure=91·8%, Precision=92%, Recall=91·7%. CIS vs RR: F-Measure=91·8%, Precision=92%, Recall=91·7%. RR vs PP: F-Measure=75·6%, Precision=75·6%, Recall=75·6%. RR vs SP: F-Measure=85·4%, Precision=85·5%, Recall=85·4%. SP vs PP: F-Measure=66·7%, Precision=67·5, Recall=65·9. CIS vs RR vs SP: F-Measure=70·6%, Precision=71·3%, Recall=70·0%	10-fold cross-validation
Kosa et al. 2016 [203]	No	Disease Progression	Supervised	CombiWISE (algorithm combines disability scoring systems)	N=408	Clinical and MRI data	.	Hold-out validation
Baranzini et al. 2015 [204]	No	Disease progression	Supervised	Random Forest	N=155	RNA biomarkers, Clinical, MRI Data	Accuracy=0·68, Sensitivity=0·22, Specificity=0·88	Hold-out validation
Wottschel et al. 2015 [205]	No	Disease Progression	Supervised	Support Vector Machine	N=74	Clinical and MRI Data	1 year follow-up: Accuracy=71·4%, Sensitivity=77%, Specificity=66%. 3 year follow up: Accuracy=68% Sensitivity=60%, Specificity=76%	Leave-one-out cross-validation
Crimi et al 2014 [206]	No	Disease Progression	Supervised and Unsupervised	Spectral clustering and Least squares linear regression	N=25	MRI Data	R^2=0·9	Leave-one-patient out cross-validation
Sweeney et al. 2014 [207]	No	Image Segmentation	Supervised	Methods Analysed: Logistic Regression, Neural Network, Support Vector Machine, Quadratic Discriminant	N=98	MRI Data	.	Hold-out validation

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
				Analysis, Linear Discriminant Analysis, Gaussian Mixture Model, k Nearest Neighbour, Random Forest, Super Learner				
Taschler et al. 2014 [208]	No	Disease Subtype	Supervised	Bayesian Spatial Generalized Linear Mixed Model or Log Guassian Cox Process	N=250	MRI Data	Bayesian Spatial Generalized Linear Mixed Model: Accuracy=0·895 (overall), 0·851 (average over all subtypes). Log Guassian Cox Process: Accuracy=0·748 (overall), 0·823 (average over all subtypes)	Leave-one-out cross-validation
Alaqtash et al. 2011 [209]	No	Diagnosis and Disease Severity	Supervised	Nearest Neighbour Classifier (k Nearest Neighbours) or Artificial Neural Network	N=20 (n(HC)=12, n(spastic diplegic cerebral palsy)=4, n(RRMS)=4)	Clinical (Ground Reaction Forces; Gait Assessment) Data	Accuracy=95%, Sensitivity=96%, Specificity=95%	Leave-one-out cross-validation
Goldstein et al. 2010 [210]	No	Risk of Disease	Supervised	Random Forest	N=3362 (n(MS)=931, n(HC)=2431)	GWAS Data	.	Out-of-bag Error
Corvol et al. 2008 [211]	No	Risk of Disease	Supervised and Unsupervised	Hierarchical Clustering and Support Vector Machine	N=62 (n(CIS)=34, n(HC)=28)	Clinical, Microarray Data	Hierarchical Clustering of high-risk group: Sensitivity=92%, Specificity=86%. Support vector machine on high-risk group: Accuracy=86%, Precision=78%, Negative Predictive Value=90%	10-fold cross-validation
Briggs et al. 2010 [212]	No	Risk of Disease	Supervised	Random Forest	N=12566 (n(test)=1343 MS + 1379 HC, n(val)=2624 MS + 7220 HC)	SNP Data	.	Independent validation dataset
Commowick et al. 2018 [213]	No	Image Segmentation	Supervised	Consensus Model	N=53	MRI Data	Dice Score~0·63, F1-score~0·5	Hold-out validation
Ohanian et al. 2016 [214]	No	Disease Classification	Supervised	Decision Tree	N=460	Questionnaire Data	Accuracy=81·2% (MS & ME or CFS), 84·0% (ME or CFS), 79·2% (MS correctly categorised)	.

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Salem et al. 2018 [215]	No	Diagnosis and Disease Monitoring	Supervised	Logistic Regression	N=60	MRI Data	Dice similarity coefficient=0·56 (segmentation), 0·77 (detection), F-score=0·806, Sensitivity=74·3%, Specificity=88·14%	Leave-one-out cross-validation
Cabezas et al. 2014 [216]	No	Disease Progression	Supervised	BOOST (ensemble classifier)	N=45 (three hospitals)	MRI Data	Median Dice Score=0·17 (hospital 1), 0·56 (hospital 2), 0·52 (hospital 3)	Leave-one-out cross-validation
Zhang et al. 2016 [217]	No	Diagnosis	Supervised	k Nearest Neighbours	N=38 and enrolled unspecified number of HCs age and gender matched	MRI Data	Accuracy=97·94%, Precision=99·09%, Sensitivity=96·15%, Specificity=99·32%	10-fold cross-validation
Birenbaum et al. 2017 [218]	No	Diagnosis and Disease Monitoring	Supervised	Convolution Neural Network	N=19 (training n=5, test n=14)	Clinical (MRI, longitudinal) Data	Cross-validation: Dice Score=0·727 Test Set: Dice Score=0·627	Leave-one-out cross-validation, independent test set
Morrison et al. 2016 [219].	No	Disease Monitoring	Supervised	Customized randomized forests and novel ensembles of randomized support vector machines	N=1041 videos	Movement Tests Data	Dice Score > 80%	.
Liu et al. 2015 [220]	No	Disease Progression	Unsupervised	Constraint-based clustering	N=266	Clinical Data	.	.
Rheumatoid Arthritis								
Chin et al. 2018 [221]	No	Risk of Disease	Supervised and Unsupervised	Non-negative Matrix Factorisation, Support Vector Machine	N=922,199 (n(RA)=1007, n(HC)=921,192)	Medical Diagnostic Database	Accuracy ~72%, Sensitivity~74%, Specificity~70%	10-fold cross-validation
Chocholova et al. 2018 [222]	No	Diagnosis and Disease Subtype	Supervised	Artificial Neural Network	N=100 (n(Seropositive RA)=31, n(Seronegative RA)=16, n(HC)=53	Immunoassay (Serum Samples) Data	Seropositive RA vs non-RA: AUC=0.96 Seronegative RA vs non-RA: AUC=0.86	Hold-out validation, testing set

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Wu et al. 2018 [223]	No	Diagnosis	Supervised	Logistic Regression	N=806 (n(HC)=383, n(T2D)=170, n(RA)=130, n(Liver Cirrhosis)=123)	Microbiome and Clinical Data	AUC=0·96, F1-score=0·92	5-fold cross-validation
Joo et al. 2017 [224]	No	Disease Progression	Supervised	Support Vector Machine	N=773 (n(train and validate)=374, n(test)=399)	GWAS & Clinical Data	Cross-validation: AUC=0·7822, Accuracy=0·7481, Sensitivity=0·7644, Specificity=0·7318. Independent Test Data: Accuracy=0·6143	10-fold cross-validation, Independent Test Data
Andreu-Perez et al. 2017 [225]	No	Disease Monitoring	Supervised	Dichotomous Mapped Forest	N=30 (n(RA)=10, n(HC)=20)	Movement Data	Accuracy 95%, F-score 81%	Leave-one-subject-out cross-validation
Orange et al. 2018 [226]	No	Disease Subtype	Both	Consensus Clustering and Support Vector Machine	N=129 (n(RA)=123, n(OA)=6)	RNA sequence and Histology Data	AUC=0·88 (high inflammatory vs other), 0·71 (low inflammatory vs other), 0·59 (mixed subtype vs other)	Leave-one-out cross-validation
Ahmed et al. 2016 [227]	No	Diagnosis	Supervised	Random Forest	N=172 (n(early OA)=46, n(early RA)=45, n(non-RA)=42, n(advanced OA)=17, n(advanced RA)=22)	Plasma amino acid analyte Data	Disease vs HC. Training set Cross-validation: AUC=0·99 Sensitivity=0·92, Specificity=0·91. Test set Cross-validation: AUC=0·96, Sensitivity=0·89, Specificity=0·9. Validation test set: AUC=0·77, Sensitivity=0·73, Specificity=0·72. Early RA classification. Training set Cross-validation: AUC=0·91, Sensitivity=0·8, Specificity=0·78. Test set Cross-validation: AUC=0·87, Sensitivity=0·77, Specificity=0·76. Validation test set: AUC=0·62, Sensitivity=0·6, Specificity=0·61.	5-fold cross-validation on training set and test set. Independent validation test set.
Miyoshi et al. 2016 [228]	No	Response to treatment	Supervised	Multilayer Perceptron	N=180	Clinical Data	AUC=0·75, Accuracy=92%, Sensitivity=96·7%, Specificity=75%	Hold-out validation

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Yeo et al. 2016 [229]	No	Early Diagnosis	Supervised	Multivariate Analysis	N=48 (n(Uninflamed Controls)=10, n(Resolving Arthritis)=9, n(early RA)=17, n(established RA)=12)	Synovial mRNA Data	Established RA vs Uninflamed: AUC=0·996 Early RA vs Resolving RA: AUC=0·764	.
Zhou et al. 2016 [570]	No	Identification of Patients	Supervised	Random Forest and C5.0 Decision Tree	N=480788	EHR Data	Test dataset 1: Accuracy=92·29% Sensitivity=86·2%, Specificity=94·6% Test dataset 2: Best-case scenario: Sensitivity=94%, Specificity=99·9%. Worst-case scenario: Sensitivity=83%, Specificity=99%	Two independent testing datasets
Lin et al. 2015 [231]	No	Identification of Patients	Supervised	Natural Language Processing and Classification Rules	N=600 (n(RA with liver toxicity)=170, n(RA)=430)	EMR Data	Cross-validation: F1-score=0·847, Precision=0·8, Recall=0·899 Test Set: F1-score=0·829, Precision=0·756, Recall=0·919	10-fold cross validation, independent test set
Chen et al. 2013 [232]	No	Identification of Patients	Supervised	Active Learning and Support Vector Machine	N=376 (n(RA)=185, n(Controls)=191)	EHR Data	AUC > 0·95	5-fold cross-validation
Lin et al. 2013 [233]	No	Disease Severity	Supervised	Natural Language Processing and Support Vector Machine	N=2017 (n(train)=852, n(test set 1)=821, n(test set 2)=344)	EMR Data	Test set 1 AUC=0·831, F1-score=0·789. Test set 2 AUC=0·785, F1 score=0·761	10-fold cross validation on two test sets
Negi et al. 2013 [234]	No	Risk of Disease	Supervised	Support Vector Machine	N=3542 (n(train)=706 RA + 761 Controls, n(test)=927 RA + 1148 Controls)	SNP Data	AUC=0·93, Accuracy=88·7%	Cross validation used
Pratt et al. 2012 [235]	No	Early Diagnosis	Supervised	Support Vector Machine	N=173 (n(RA)= 47, n(non-RA)=64, n(undifferentiated arthritis)=62)	CD4 T Cell Transcriptome Data	Sensitivity=0·68, Specificity=0·7. Removing ACPA-positive subset: Sensitivity=0·85, Specificity=0·75	Hold out validation
Singh et al. 2012 [236]	No	Diagnosis	Supervised	Fuzzy Inference System	N=150	Clinical Data	.	.

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Kruppa et al. 2012 [237]	No	Risk of Disease	Supervised	Random Forest in regression mode (Random Jungle)	N=1445 (n(RA)=707 and n(HC)=738)	GWAS Data	AUC=0·8925	Hold-out validation
Liu et al. 2011 [238]	No	Risk of Disease	Supervised	Random Forest	N=4880 (n(cohort 1)=908 RA + 1260 controls, n(cohort 2)= 952 RA + 1760 controls)	SNP Data	Accuracy=70%, Sensitivity=74%, Specificity=66%	Out of bag error, Independent validation cohort
Nair et al. 2010 [239]	No	Response to treatment	Supervised	Least Squares Kernel-Conjugate gradient algorithm	N=25 (n(RA)=8, n(OA)=10, n(HC)=7)	Electro-myographic Gait Data	Accuracy=91·07%, Sensitivity=81%, Specificity=82%	8-fold cross-validation
Briggs et al. 2010 [240]	No	Risk of Disease	Supervised	Random Forest and Logistic Regression	N= 4130	SNP Data	.	Hold-out validation
Niu et al. 2010 [241]	No	Diagnosis	Supervised	Boosted Decision Tree	N=143 (n(RA)=43, n(AID Controls)=50, n(HC)=50)	Mass Spectrometry (from serum)	Accuracy=85·7% (RA), 87·5% (autoimmune controls), 88·0% (HC). Sensitivity=85·71%, Specificity=87·76% (RA vs controls)	Hold-out validation
Geurts et al. 2005 [242]	Yes	Diagnosis	Supervised	Decision Trees (RA Boosting, IBD Extra-Trees)	N(RA)=206 (68 RA, 138 controls), N(IBD)=480 (240 IBD, 240 controls)	Mass Spectrometry (from serum)	RA: Sensitivity=83·82%, Specificity=94·93% IBD: Sensitivity=88·33%, Specificity= 91·63%	Leave-one-out cross-validation
de Seny et al. 2005 [243]	Yes	Early Diagnosis	Supervised	Decision Tree Boosting	N=103 (n(RA)=34, n(inflammatory controls)=20 PsA + 9 Asthma + 10 CD, n(controls)=14 OA + 16 HC)	Mass Spectrometry (from serum)	RA vs controls: Sensitivity=85%, Specificity=91% (2 independent spectra), Sensitivity=94%, Specificity=90% (2 combined spectra). RA vs PsA: Sensitivity=94%, Specificity=86% (2 independent spectra), Sensitivity=97%, Specificity=76% (2 combined spectra).	Leave-one-out cross validation
Scheel et al. 2003 [244]	No	Early Diagnosis	Supervised	Neural Network, Method in [571]	N=22 patients, N=72 joints examined	Laser Imaging Data	Accuracy=86%, Sensitivity=80%, Specificity=89%	.
Gronsbell et al. 2018 [245]	No	Identification of Patients	Supervised and Unsupervised	Unsupervised (clustering based) Feature Selection and Sparse Regression	N=435	EMR Data	AUC=0·928	Independent validation dataset
Gossec et al. 2018 [246]	Yes	Disease Monitoring	Supervised	Multiclass Selective Naïve Bayes Classifier	N=155 (82 RA, 73 axSpA)	Physical Activity Data	Sensitivity=95·7%, Specificity=96·7%	Hold-out validation

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Lezcano-Valverde et al. 2017 [247]	No	Mortality	Supervised	Random Survival Forests	N=1741	Demographic & Clinical Data	1 year follow-up: Sensitivity=0·79, Specificity=0·8. 7 year follow up: Sensitivity=0·43, Specificity=0·48.	Hold-out validation
Gonzalez-Recio et al. 2009 [248]	No	Risk of Disease	Supervised	Information gain/entropy reduction criteria and Bayesian threshold LASSO	N=2062 (n(cases)=868, n(controls)=1194)	SNPs	.	5-fold cross-validation
Heard et al. 2014 [249]	No	Early Diagnosis	Supervised	Artificial Neural Network and Decision Tree	ANN: N=300 (n(HC)=98 n(OA)=101, n(RA)=101) DT: N=298 (n(HC)= 100, n(OA)=100, n(RA)=98)	Clinical (Inflammatory cytokine expression, serum samples) Data	ANN: Sensitivity=100% (HC), 100% (OA), 100% (RA), Specificity=100% (HC), 100% (OA), 100% (RA) for all cytokines and significant cytokines. DT: Sensitivity=100% (HC), 100% (OA), 95% (RA), Specificity=96% (HC), 97% (OA), 100% (RA) for all cytokines.	Hold-out validation, independent testing set
Gronsbell et al. 2018 [250]	Yes	Identification of Patients	Semi-Supervised	Semi-supervised approach	N(RA)=44014 (500 labelled, 43514 unlabelled), N(MS)=12198 (455 labelled, 11743 unlabelled)	EMR Data	AUC=94·93 (RA), 93·94 (MS)	10-fold cross-validation
Van Looy et al. 2006 [251]	No	Response to treatment	Supervised	Multilayer Perceptron or Support Vector Machine	N=511	Clinical Data	All Cases: AUC=0·772, Sensitivity=0·95, Specificity=0.402 or Sensitivity=0·265, Specificity=0·95 (MLP). Complete Cases, MLP: AUC=0·854, Sensitivity=0.95, Specificity=0·548 or Sensitivity=0·462, Specificity=0·95. Complete Cases, SVM: AUC=0·863, Sensitivity=0·95, Specificity=0·507 or Sensitivity=0·308, Specificity=0·95. Expectation Maximisation, MLP: AUC=0·813, Sensitivity=0·95, Specificity=0·411, or Sensitivity=0·412, Specificity=0·95. Expectation Maximisation, SVM: AUC=0·804, Sensitivity=0·95, Specificity=0·402, or Sensitivity=0·412, Specificity=0·95.	.

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Wyns et al. 2004 [252]	No	Early Diagnosis	Supervised and Unsupervised	Kohonen Neural Network (includes Self Organising Maps)	N=160 (n(RA)=51 RA, n(SpA)=43, n(other)=26, n=40 with no definite diagnosis)	Clinical Data	Accuracy=62·3%, 65·3% (without undetermined samples)	Hold-out validation
Inflammatory Bowel Disease								
Waljee et al. 2018 [253]	No	Disease Progression	Supervised	Random Forest	N =20368	Clinical Data	Predict Hospitalisation and Corticosteroid Prescriptions. IBD: AUC=0·87, Sensitivity=74-80%, Specificity=80-82%. UC: AUC=0·84. CD=0·85. IC=0·82. Predict Corticosteroid Prescription Only, IBD: AUC=0·9 Predict Hospitalisation and Corticosteroid Prescriptions (12 month outcome): AUC=0·9	Hold-out validation
Mossotto et al. 2017 [254]	No	Disease Subtype	Supervised and Unsupervised	Support Vector Machine, Hierarchical Clustering	N=287 Training and testing: N=210 (n(CD)=178, n(UC)=80, n(IBDU)=29 (only reclassified))	Clinical Data	Cross-validation: AUC=0·87, Accuracy=82·7%, Precision=0·91, Recall=0·83, F1-score=0·87. Independent test set: Accuracy=83·3%, Precision=0·86, Recall=0·83, F1-score=0·84	5-fold cross validation, independent test set
Maeda et al. 2018 [255]	No	Disease Severity	Supervised	Support Vector Machine	N=187	Endocytoscopic Image Data	Accuracy=91%, Kappa=1, Sensitivity=74%, Specificity=97%	Hold-out validation
Douglas et al. 2018 [572]	No	Diagnosis and Response to Treatment	Supervised	Random Forest	N=771 (n (test)=40 (n(CD)=20, n(HC)=20). n(validation, diagnosis only) = 731 (444 CD, 287 control))	Metagenomic Data	Diagnosis: Accuracy=84·2%. Independent Validation (diagnosis): Accuracy 73·2%. Treatment Response: Accuracy 77·8%.	Out of bag error, Leave-one-out cross-validation, Independent test data (diagnosis only)
Jain et al. 2017 [573]	No	Disease Progression	Supervised	Decision Tree	N=179	Clinical Data	Colectomy Prediction: Accuracy=77%, Sensitivity=75%, Specificity=80%. Steroid Dependence: Accuracy=75%, Sensitivity=69%, Specificity=80%.	Hold-out Validation

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Waljee et al. 2017 [258]	No	Response to Treatment	Supervised	Random Forest	N=1080	Clinical Data	Objective Remission: AUC=0·79, Sensitivity=70·6%, Specificity=73·8%. Non-adherence: AUC=0·84, Sensitivity=70·6%, Specificity=85·0%. Shunting: AUC=0·78, Sensitivity=65·2%, Specificity=79·0%.	Out of bag error, Hold-out validation
Isakov et al. 2017 [259]	No	Risk of Disease	Supervised	Combined Model (elastic net regularised generalised linear model, extreme gradient boosting, support vector machine, random forest)	N = 513 (n(CD)=180, n(UC)=149, n(colorectal neoplasms)=94, n(normal tissue)=90)	Gene Expression Data	AUC=0·829, Accuracy=0·808, Sensitivity=0·577, Specificity=0·880	5-fold cross-validation
Kang et al. 2017 [260]	No	Response to Treatment	Supervised	Gene Regulatory Network-based Regularized Artificial Neural Network (GRRANN)	N=46	Gene Expression Data	Balanced Accuracy≈0·8	5-fold cross validation, Hold-out validation
Waljee et al. 2018 [261]	No	Response to Treatment	Supervised	Random Forest	N=491	Clinical Data	AUC=0·73, Sensitivity=0·72, Specificity=0·68	Hold-out validation
Pal et al. 2017 [282]	No	Risk of Disease	Supervised	Consensus Method (Naïve Bayes, Logistic Regression, Random Forest)	N=111 (n(CD)=64, n(HC)=47)	GWAS Data, Exome Data to impute genotypes.	AUC=0·72	Hold-out validation
Eck et al. 2017 [262]	No	Diagnosis	Supervised	Support Vector Machine or Random Forest	N=112 (n(IBD)=56, n(HC)=56)	Microbiota Data	Accuracy=81%	10-fold cross validation
Menti et al. 2016 [263]	No	Disease Progression	Supervised	Bayesian Networks	N=152	Clinical Data and Selected Genetic Data	AUC=0·95, Accuracy=0·89, Sensitivity=0·78, Specificity=0·94	10-fold cross validation

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Hubenthal et al. 2015 [264]	No	Diagnosis	Supervised	Support Vector Machine	N=273 (n(CD)=37, n(UC)=32, n(HC)=92, n(COPD)=23, n(MS)=23, n(pancreatitis)=35, n(sarcoidosis)=32)	MicroRNA Expression Data	AUC=0·95, Balanced Accuracy=0·95, Sensitivity=1, Specificity=0·9	5-fold cross-validation
Niehaus et al. 2015 [265]	No	Disease Severity	Supervised and Unsupervised	Support Vector Machine, Hierarchical Clustering	N=501	Health Records, EMR Databases	Accuracy=68·7%, Sensitivity=59·1%, Specificity=78·4%	5-fold cross validation, testing dataset
Wei et al. 2013 [266]	No	Risk of Disease	Supervised	Logistic Regression	N=53,279 (n(CD)=17,379, n(UC)=13,458, n(HC)= 22,442	GWAS Data	Cross Validation: AUC=0·864 (CD) 0·83 (UC). Independent Test Set: AUC=0·864 (CD), 0·826 (UC)	10-fold cross validation, independent testing dataset
Cui et al. 2013 [267]	No	Diagnosis	Supervised	Support Vector Machine	N=124 (n(IBD)=25, n(HC)=99)	Metagenomic Data	Accuracy=88%, Sensitivity=92%, Specificity=84%	Leave-one-out cross-validation
Waljee et al. 2010 [268]	No	Response to Treatment	Supervised	Random Forest	N=346	Clinical Data	AUC=0·856 (response), 0·813 (non-adherence), 0·797 (shunting)	10-fold cross validation, validation data set
Firouzi et al 2007 [269]	No	Disease Progression	Supervised	Decision Tree	N=160 (121 UC, 39 CD)	Clinical Data	Accuracy=88·2% (UC), 89·8% (CD), 86·5% (IBD), Sensitivity=67·6% (UC), 82·8% (CD), 65·7% (IBD), Specificity=96·3% (UC), 95·2% (CD), 96·3% (IBD), Matthew's Correlation Coefficients=0·69 (UC), 0·79 (CD), 0·68 (IBD)	10-fold cross-validation
Ozawa et al. 2018 [270]	No	Disease Severity	Supervised	Neural Network	N= 30,285 images, N=558 patients	Colonoscopy White-light Image Data	Mayo 0 vs Mayo 1-3: AUC=0·86. Mayo 0-1 vs Mayo 2-3: AUC=0·98	Hold-out validation
Reddy et al. 2018 [271]	No	Disease Severity	Supervised	Gradient Boosting Machines	N=82	EHR Data	AUC=92·82%	10-fold cross-validation
Forbes et al. 2018 [272]	Yes	Diagnosis	Supervised	Random Forest	N=102 (n(CD)=20, n(UC)=19, n(MS)=19, n(RA)=21, n(HC)=23)	Microbiota Data	Diseased vs HC: AUC=0·93, Balanced Accuracy=0·84. Breakdown per inflammatory disease found in paper	Out of bag error

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Doherty et al. 2018 [273]	No	Response to treatment	Supervised	Random Forest	N=306 (n(CD treated)=232, n(CD untreated)=74)	Microbial Genome Data and Clinical Data	Remission: AUC=0·844, Sensitivity=0·774, Specificity=0·831. Response: AUC=0·733 Sensitivity=0·684, Specificity=0·724	.
Han et al. 2018 [274]	No	Disease subtype	Supervised	Random Forest	N=163 (n(train)=24 CD, 59 UC, n(Validation set 1)=5 CD , 7 UC, n(Validation set 2)=14 CD, 10 UC, n(Validation set 3)=11 CD, 5 UC, n(Validation set 4)=13 CD, 15 UC) Biopsy Samples	Gene Expression Data	Validation set 1: AUC=0·829 Validation set 2: AUC=0·764 Validation set 3: AUC=0·836 Validation set 4: AUC=0·849	Hold-out validation
Daneshjou et al. 2017 [275]	No	Risk of Disease	Supervised	Metaclassifier	N=111 (n(CD)=64, n(HC)=47)	Exome-sequencing data	AUC=0·78	Cross-validation performed
Giollo et al. 2017 [276]	No	Risk of Disease	Supervised	Support Vector Machine or Ensemble Classifier	N=111 (n(cases)=64, n(controls)=47)	Genetic Data	AUC=0·6 (SVM), 0·66 (Ensemble Classifier)	Cross validation performed
Yu et al. 2017 [277]	Yes	Identification of Patients	Supervised	Natural Language Processing	N= 2393 (435 RA, 758 CAD, 600 UC, 600 CD)	Electronic Medical Records Data	AUC~0·94 (RA), ~0·95 (CD), ~0·95 (UC) F-score ~0·71 (RA), ~0·83 (CD), ~0·89 (UC)	.
Wisittipanit et al. 2015 [278]	No	Diagnosis	Supervised	Support Vector Machine	N=425 (n(CD)=101, n(UC)=89, n(HC)=235 HC)	LH-PCR (Microbiome) Data	AUC=0·73 (CD), 0·78 (UC), 0·77 (HC), Accuracy=78·18% (CD), 79·71% (UC), 75·62% (HC)	5-fold cross validation
Ahmed et al. 2017 [279]	No	Diagnosis	Supervised	Neuro-Fuzzy Automated Classifier	N=387 (n(CD)=144, n(HC)=243)	Genetic Data	Accuracy=97·67%, Sensitivity=96·07%, Specificity=100%	Hold-out validation, testing set
Mahapatra et al. 2016 [280]	No	Image Segmentation	Semi-Supervised	Random Forest-based Classifier	N=70 (CD)	MRI Data	Dice metric=92·4%, Hausdorff=7mm	5-fold cross validation

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Mahapatra et al. 2016 [281]	No	Image Segmentation	Supervised	Random Forest	N=50 (CD)	MRI Data	Dice metric=91·7%, Hausdorff=7.4mm	5-fold cross validation
Type 1 Diabetes								
Stawiski et al. 2018 [283]	No	Diagnosis	Supervised	Artificial Neural Network	N=315	Clinical Data	R^2=0·6455	Hold-out validation
Ben Ali et al. 2018 [284]	No	Disease Management	Supervised	Artificial Neural Network	N=12 patients, N=1344 samples	CGM Data	Average RMSE=6·43 (mg/dL)	Hold-out validation
Perez-Gandia et al. 2018 [299]	No	Disease Management	Supervised	Decision Support System with Artificial Neural Network	N= 21 patients, longitudinal analysis	Clinical Data	.	Hold-out validation
Maulucci et al. 2017 [285]	No	Diagnosis and Disease Monitoring	Supervised	Decision Support System	N=26	RBC Image Data	Control): Accuracy=1, Precision=1, Recall=1, F1-score=1. T1D: Accuracy=1, Precision=1, Recall=1, F1-score=1. T1D with complications: Accuracy=1, Precision=1, Recall=1, F1-score=1.	Leave-one-person-out cross-validation
Siegel et al. 2017 [286]	No	Disease Management	Supervised	Linear Discriminant Analysis	N=52 patients, N=128 samples.	VOCs	AUC=0·895, Sensitivity=91%, Specificity=84%	Leave-one-out cross-validation
Zhao et al. 2016 [287]	No	Risk of Disease	Supervised	LASSO (regression)/OOR (developed method)	N=1418 (n(T1D)=962 T1D, n(controls)= 448	Genetic Data	AUC=0·89	Hold-out validation
Georga et al. 2015 [288]	No	Disease Management	Supervised	KOS-ELM (online sequential extreme learning machine kernels)	N=15, longitudinal analysis	Clinical Data	Case 1: RMSE=16·6 (mg/dl) Case 2: RMSE=10·9 (mg/dl) Case 3: RMSE=8·5 (mg/dl)	10-fold cross validation
Georga et al. 2013 [289]	No	Disease Management	Supervised	Support Vector Regression	N=15 patients, longitudinal analysis	Clinical Data	Nocturnal: Sensitivity=0·94, Precision=0·98 (30 minutes and 60 minutes).	10-fold cross validation

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
							Diurnal: Sensitivity=0·92, Precision=0·93 (30 minutes), Sensitivity=0·96, Precision=0·97 (60 minutes)	
Marling et al. 2013 [290]	No	Disease Management	Supervised	Support Vector Machine Regression	N=19 patients, N=262 CGM plots	CGM Data	Accuracy=90·1%, Sensitivity=97%, Specificity=74·1%	10-fold cross-validation
Nguyen et al. 2013 [291]	No	Risk of Disease	Supervised	RIPPER (decision rules) and Logistic Regression Method (Predict DQ types without DR type information)	N=10579 (n(train)=7405, n(test)=3174)	SNP Data	Independent Test Dataset. Predict HLA Types: AUC=0·997 Accuracy=99·3%. Predict High Risk HLA types AUC=0·995 Accuracy=99·8%. Predict high risk subtype (DRB1*03:01-DQA1*05:01-DQB1*02:01): AUC=0·998, Accuracy=99·8%. Predict DQ Types without DR type information: AUC=0·98.	10-fold cross validation, independent test dataset
Wei et al. 2009 [292]	No	Risk of Disease	Supervised	Support Vector Machine	N=8438 (n(WTCCC-T1D)=1963 cases + 1480 controls ,n(CHOP/Montreal-T1D)= 1008 cases + 1000 controls, n(GoKinD-T1D)=1529 cases + 1458 controls)	GWAS Data	WTCCC-T1D dataset: AUC=0·89, Sensitivity=0·87, Specificity=0·75. CHOP/Montreal-T1D dataset: AUC=0·83, GoKinD-T1D dataset: AUC=0·84	5-fold cross-validation
Jensen et al. 2014 [293]	No	Disease Management	Unsupervised	Pattern Classification Algorithm	N=10 patients, longitudinal measurements (20 x sessions with Professional CGM)	CGM Data	Sensitivity=78%, Specificity=96%, (All hypoglycaemic events detected, 1 false positive)	.
Schwartz et al. 2008 [294]	No	Disease Management	Supervised	Case-based reasoning	N=12 patients, longitudinal measurements	Clinical Data	.	.
Cordelli et al. 2018 [295]	No	Diagnosis and Diseases Monitoring	Supervised	Support Vector Machine	N=27 (n(HC)=8, n(T1D)=10, n(T1D with complications)=9	RBC Images	F1 score=1, Precision=1, Recall=1 (for HC, T1D, and T1D with complications)	Leave-one-person-out cross-validation

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Sampath et al. 2016 [296]	No	Disease Management	Supervised	Aggregating ranking algorithms in machine learning	N=213 (n(DIAdvisor)=34, n(ChildrenData)=179)	Clinical Data	Sensitivity=77·03%, Specificity=83·46%	Independent validation dataset
Georga et al. 2015 [297]	No	Disease Management	Supervised	Random Forest (feature selection), Support Vector Regression or Gaussian processes	N=15 patients, longitudinal measurements	Clinical Data	30min prediction horizon: SVR RMSE=5·7, GP RMSE=5·6; 60min prediction horizon: SVR RMSE=6·4, GP RMSE=6·3	10-fold cross-validation
Ling et al. 2016 [298]	No	Disease Management	Supervised	Extreme learning machine-based neural network	N=16 patients, N=589 samples	Clinical Data	Gamma value=70·8%, Sensitivity=78%, Specificity=60%	Noted by researchers that cross-validation is not required.
Systemic Lupus Erythematosus								
Ceccarelli et al. 2018 [300]	No	Disease Progression	Supervised	Logistic Regression	N=120	Clinical Data	AUC=0·806	Leave-one-out cross-validation
Turner et al. 2017 [301]	No	Identification of Patients	Supervised	Natural Language Processing and Neural Network or Random Forest	N=662 (n(SLE)=332, n(HC)=340)	EHR Data	AUC=0·974 (Neural Network), 0·988 (RF), Accuracy=92·1% (Neural Network), 95% (Random Forest)	5-fold cross-validation
Ceccarelli et al. 2017 [302]	No	Disease Progression	Supervised	Recurrent Neural Networks	N=132 (n(develop chronic damage)=38, n(no chronic damage)=94)	Clinical Data	AUC=0·77, Sensitivity=0·74, Specificity=0·76	8-fold cross-validation
Kan et al. 2016 [303]	No	Disease Progression	Unsupervised	Cluster Analysis	N=1611	Demographic & Drug Treatment	.	Cross-validation not recommended for cluster analysis
Wolf et al. 2016 [304]	No	Treatment Response	Supervised	Random Forest	N=140 (n(non-responders)=103, n(responders)=37)	Urine Biomarkers	AUC=0·79, Sensitivity=0·76, Specificity=0·73	Cross-validation not required for Random Forest
Guy et al. 2012 [305]	No	Risk of Disease	Supervised	Bagged Alternating Decision Trees	N=6728 (1846 SLE + 1825 Controls)	SNPs	.	.

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Tang et al. 2011 [306]	No	Mortality	Supervised	Logistic Regression	N= 3313	Clinical Record Data	AUC=0·74	10-fold cross-validation
Armananzas et al. 2009 [307]	Yes	Diagnosis	Supervised and Unsupervised	Consensus Method	N=14 (n(HC)=6, n(SLE)=3, n(PAPS)=5)	Microarray Expression Data	.	10-fold cross-validation
Huang et al. 2009 [308]	No	Diagnosis	Supervised	Decision Tree	N=232 (n(SLE)=64, n(AID controls)=85, n(HC)=83)	Serum Proteome Data	SLE: Accuracy=78·1%, Sensitivity=78·1%, Specificity=96·3% AID Controls: Accuracy=85·8%, Sensitivity=85·7%, Specificity=86·7% HC: Accuracy=90%, Sensitivity=90%, Specificity=96%.	Hold-out validation
Murray et al. 2018 [309]	No	Identification of Patients	Supervised	Logistic Regression	N=17057 (n(SLE)=583, n(control)=16174, n(potential SLE)=150, n(random)=150)	EHR Data	AUC=0·97, Accuracy=0·92, Precision=0·85, Recall=0·97	Hold-out validation
Reddy et al. 2018 [310]	No	Disease Progression	Supervised	Recurrent Neural Network	N=9457	EHR Data	AUC=0·7, Accuracy=70·54%, Sensitivity=74·49%, Specificity=56·61%	Hold-out validation
Tang et al. 2018 [311]	No	Disease Progression	Supervised	Random Forest and Multilinear Regression	N=173	Clinical Data	Random Forest, multi-classifier: Accuracy=53·7% (Class II), 56·2% (Class III&IV):56·2%, 40·1% (Class V). Random Forest, binary classifier: Accuracy=56·2% (Class II), 63·7% (Class III&IV), 61% (Class V). Multilinear regression: CI prediction: Q^2=0·746, R^2=0·771. AI prediction: Q^2=0·516, R^2=0·576.	5-fold cross validation (Predicting AI and CI)
Scully et al. 2010 [312]	No	Diagnosis	Supervised	Naïve Bayesian Classifier and Support Vector Machine	N=27	MRI Data	Leave-one-out training data: Sensitivity=94·3%, Specificity= 93·1% Test data: Sensitivity=94·3%, Specificity=93.9%	Leave one out cross validation, Test dataset
Davis et al. 2013 [313]	No	Risk of Disease	Supervised	Random Jungle, ReliefF or evaporative cooling	N=404 (n(SLE)=209, n(HC)=195)	Exome Data	.	.

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Psoriasis and Psoriatic Arthritis								
Wang et al. 2016 [314]	No	Diagnosis	Supervised	Random Bits Forest (Neural Network, Boosting, Random Forest)	N=2723 (n(train)=915 cases + 675 controls; n(test)=431 cases + 702 controls)	GWAS Data	Cross-validation: AUC=0·6739, Accuracy=0·639, Sensitivity=0·6317, Specificity=0·649. Test Dataset: AUC=0·7239, Accuracy=0·692, Sensitivity=0·6543, Specificity=0·7151.	10-fold cross validation, independent testing dataset
George et al. 2018 [315]	No	Disease Severity	Supervised and Unsupervised	Unsupervised Feature Learning, Random Forest	N=676 images, N=44 patients	Digital Image Data	F1-score=0·71	10-fold cross validation
Shrivastava et al. 2017 [316]	No	Disease Severity	Supervised	Support Vector Machine	N=670 images, N=110 patients	Digital Image Data	AUC=0·998, Accuracy=99·84%, Sensitivity=99·76%, Specificity=99·99%	10-fold cross validation
Shrivastava et al. 2016 [317]	No	Diagnosis	Supervised	Support Vector Machine	N=540 (n(HC)=270, n(P)=270) images, N=30 patients.	Digital Image Data	AUC=1, Accuracy=100%, Sensitivity=100%, Specificity=100%	10-fold cross validation
Shrivastava et al. 2016 [318]	No	Disease Severity	Supervised	Support Vector Machine	N=848 images, N=65 patients	Digital Image Data	Accuracy=99·92%	10-fold cross validation
Shrivastava et al. 2015 [319]	No	Diagnosis	Supervised	Support Vector Machine	N=540 (n(HC)=270, n(P)=270) images, N=30 patients.	Digital Image Data	AUC=0·999, Accuracy=99·94%, Sensitivity=99·93, Specificity=99·96%	10-fold cross validation
Cowen et al. 2007 [320]	No	Diagnosis	Supervised	Partial Least Squares Regression, Support Vector Machine and C5.0 Decision Tree	N=148 (n(tumour-stage MF)=45, n(psoriasis)=56, n(HC)=47)	Proteomic Data from Serum	Tumour-Stage MF vs Psoriasis: Sensitivity=78·57%, Specificity=93·75% (Ciphergen), Sensitivity=78·57, Specificity=86·67% (PrOTOF). Psoriasis vs HC: Sensitivity=93·75%, Specificity=75% (Ciphergen), Sensitivity=86·67%, Specificity=76·92%. (PrOTOF).	10-fold cross validation, independent testing dataset
Raina et al. 2016 [321]	No	Disease Severity	Supervised	Linear Discriminant Analysis	N=20 patients, N=80 images	Digital Image Data	Accuracy=48·75%, Kappa=0·4203	Leave-one-out cross-validation
Shrivastava et al. 2015 [322]	No	Diagnosis	Supervised	Support Vector Machine	N=540 (n(HC)=270, n(P)=270) images, N=30 patients.	Digital Image Data	AUC=1, Accuracy=99·81%, Sensitivity=99·26%, Specificity=97·04%	Jack Knife (N fold) cross-validation

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Shrivastava et al. 2016 [323]	No	Diagnosis	Supervised	Support Vector Machine	N=540 (n(HC)=270, n(P)=270) images, N=30 patients.	Digital Image Data	AUC=0·99, Accuracy=99·39%, Sensitivity=99·43%, Specificity=99·35%	10-fold cross-validation
Patrick et al. 2018 [324]	Yes	Risk of Disease and Disease Progression	Supervised	Conditional Inference Forest or Shrinkage Discriminant Analysis	N=22181 (n(PsV)=7855, n(PsA)=2703, n(PsC)=2681, n(HC)=8942)	GWAS Data	AUC=0·82 (cross validation and holdout test set)	Cross-validation performed, test set
Coeliac Disease								
Hujoel et al. 2018 [325]	No	Diagnosis	Supervised	Random Forest or Bagged Classification Trees	N = 408	EMR Data	AUC≈0·55	10-fold cross-validation
Arasaradnam et al. 2014 [326]	No	Diagnosis	Supervised	Logistic Regression	N=47 (n(D-IBS)=20, n(CeD)=27)	VOCs Data	AUC=0·91, Sensitivity=85%, Specificity=85%	Leave-one-out cross-validation
Tenorio et al. 2011 [327]	No	Diagnosis	Supervised	Bayesian Classifier (Average One-Dependence Estimator)	N=216 (CeD 46% of records in training data, 37% in test data)	Clinical Data	AUC=0·84, Accuracy=80%, Sensitivity=0·78, Specificity=0·80	10-fold cross-validation
Choung et al. 2018 [328]	No	Diagnosis and Disease Monitoring	Supervised	Random Forest (peptide selection), Support Vector Machine	Diagnosis: N= 468 (n(CeD)= 172, n(HC)=296). Monitoring: N= 465 (n(CeD treated, healed)=85, n(CeD treated, unhealed)=81, n(CeD, untreated)=82, n(HC)=217, n(disease controls)=27).	Peptide Data	Diagnosis: Accuracy=99%, Sensitivity=99%, Specificity=100%. Monitoring: Accuracy=90%, Sensitivity=84%, Specificity=95%	Hold-out validation (diagnosis only)
Chen et al. 2016 [329]	No	Diagnosis	Supervised	Logistic Model	N=1498 (n(CeD)=363, n(FP)=1135)	EHR Data	AUC=0·94, F1-score=0·92, Kappa=0·78, Precision=0·93, Recall=0·92	10-fold cross-validation
Ludvigsson et al. 2013 [330]	No	Diagnosis	Supervised	Natural Language Processing	N=496 (n(train)=327, n(test)=169)	EMR Data	F-measure 84·5%, Sensitivity=72·9%, Specificity=89·9%	Hold-out validation

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Amirkhani et al. 2018 [331]	No	Disease Severity	Supervised	Combined fuzzy cognitive map and possibilistic fuzzy c-means clustering algorithm	N=89	Clinical Data	Accuracy=91% (A), 90% (B1), 88% (B2)	Leave-one-out cross-validation
Thyroid Disease								
Ahmad et al. 2018 [332]	No	Diagnosis	Supervised	Hybrid model (linear discriminant analysis, k-nearest neighbour weighed preprocessing, adaptive neurofuzzy inference system)	N=3163 (n(hypo)=152, n(negative)=3011)	Clinical Data	Accuracy=98·5, Sensitivity=94·7%, Specificity=99·7%	10-fold cross validation
Baccour L. et al 2018 [333]	No	Diagnosis	Supervised	ATOVIC (hybrid multi-criteria decision making method)	N=7200	Clinical Data	Accuracy=92·7%, F-measure=95·3% (Hyper- vs Hypo- vs Control). Accuracy=99·81% (Hypo- vs Control)	Hold-out validation
Morejon et al. 2017 [334]	No	Diagnosis	Supervised	Java Agent Framework for Health Data Mining	.	Clinical Data	.	Hold-out validation
Temurtas et al. 2009 [335]	No	Diagnosis	Supervised	Probabilistic Neural Network	N=215 (n(normal)=150, n(hypo)=30, n(hyper)=35)	Clinical Data	Accuracy=94·81%	10-fold cross validation
Polat et al. 2007 [336]	No	Diagnosis	Supervised	Artificial Immune Recognition System with fuzzy weighted pre-processing	N=215 (n(normal)=150, n(hypo)=30, n(hyper)=35)	Clinical Data	Accuracy=85%	10-fold cross validation
Keles et al. 2008 [337]	No	Diagnosis	Supervised	Expert system for thyroid disease diagnosis with fuzzy rules	N=215 (n(normal)=150, n(hypo)=30, n(hyper)=35)	Clinical Data	Accuracy=95·33%	10-fold cross validation
Autoimmune Liver Disease								
Weiss J et al. 2015 [338]	No	Response to Treatment	Supervised	Boosted Forest	N=288	Clinical Trial Data	.	Hold-out validation

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Singh et al. 2017 [339]	No	Disease Progression	Supervised	Kullback-Leibler Divergence-Least Squares Support Vector Machine	N=276	Clinical Data	Accuracy=90·94%	Hold-out validation
Eaton et al. 2018 [340]	No	Disease Progression	Supervised	Gradient Boosting	N=787	Clinical Data	Cross-validation: C-statistic=0·96 Independent test data: C-statistic=0·9	5-fold cross validation, independent test dataset
Iwasawa et al. 2018 [341]	Yes	Diagnosis	Supervised	Random Forest	N= 64 (n(PSC)=24, n(UC)=16, n(HC)=24)	Microbiome Data	Genera: AUC=0·7423 (PSC vs HC), 0·8756 (PSC vs UC). Species: AUC=0·8756 (PSC vs HC), 0·7626 (PSC vs UC)	10-fold cross-validation
Tsujitani et al. 2009 [342]	No	Survival Prediction	Supervised	Neural Network	N=312	Clinical Data	.	Delete-one cross-validation
Systemic Sclerosis								
Zhu et al 2018 [343]	No	Diagnosis	Supervised and Unsupervised	Hierarchical Clustering and Support Vector Machine	N=37 (n(controls)=19, n(SSc)=18)	DNA and RNA of PBMC	Accuracy=100%, Sensitivity=100%, Specificity=100%	Hold-on-one-out cross-validation
Taroni et al. 2017 [344]	No	Response to treatment	Supervised	Support Vector Machine	.	Gene expression Data	.	.
Huang et al. 2015 [345]	No	Disease Progression	Supervised	Random Forest	N=119	Clinical and peripheral blood flow cytometry Data	Accuracy=95%	Hold-out cross-validation
Berks et al. 2014 [346]	No	Diagnosis	Supervised	Random Forest	N= 991 (n(train)=80 ; n(validate)=104 HC + 83 PR + 269 SSc; n(test)=104 HC + 83 PR + 268 SSc) images	Nailfold Capillaroscopy Data	Accuracy=93·6%, F-measure=71·5%, Precision=64·1%, Recall=80·9%	Hold-out validation, testing set

Paper	Multiple AIDs Studied	Prediction or Classification Task	ML Type	Machine Learning Method	Study Size (N)	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set.	Cross-Validation
Alopecia								
Huang et al. 2013 [347]	Yes	Comorbidity analysis	Supervised	Natural Language Processing	N=3568 (n(AA)=2115) and N=416 (PAFS cohort)	Patient Data Repository	Validity=93·9%	Hold-out validation
Vitiligo								
Sheth et al. 2013 [348]	Yes	Comorbidity analysis	Supervised	Natural Language Processing	N=3280	Research Patient Data Repository	.	.

Supplementary Table 3 Pathways excluded from Section 5.3.2's optimal IBD subtype classifier that utilises the AI gene panel. Pathways listed here were enriched in the Enrichr pathway analysis of the AI panel genes after pre-processing prior to feature selection, but were **not** enriched in the Enrichr pathway analysis of the gene selected during feature selection for the IBD subtype ML model.

Term	**Overlap in AI Panel**	**P-value**	**Adjusted P-value**	**Odds Ratio**	**Combined Score**	**Genes**
Mitophagy	21/68	2.00E-08	4.40E-08	5.42	96.01	*PRKN, JUN, SRC, RRAS2, FOXO3, HIF1A, RELA, MAPK10, MAPK9, MAPK8, NRAS, TBK1, SP1, TAX1BP1, E2F1, KRAS, SQSTM1, TP53, ATF4, ATG5, BCL2L1*
Cholinergic synapse	28/113	2.20E-08	4.80E-08	4.00	70.58	*CAMK2D, PIK3CD, ADCY3, ITPR3, PIK3R2, PIK3R1, ADCY7, GNAI2, NRAS, AKT2, CREB3L2, MAPK1, FYN, PRKACA, JAK2, CAMK2G, KCNJ2, MAPK3, PRKCB, FOS, CREB3, CREB1, PIK3CA, BCL2, KRAS, PLCB1, PLCB2, ATF4*
Hippo signaling pathway	35/163	2.23E-08	4.83E-08	3.33	58.68	*GSK3B, YWHAB, SERPINE1, TCF7, ITGB2, LEF1, PPP2CB, CCND3, PAK1, CCND1, MYC, DVL2, YWHAH, SMAD2, SMAD1, WNT10B, TCF7L2, SMAD4, TGFB2, SMAD3, TGFB1, FBXW11, TGFB3, BMP8A, CSNK1D, TGFBR1, SMAD7, TGFBR2, APC, PARD3, ID1, BIRC5, CTNNB1, BIRC2, BIRC3*
Endocytosis	44/252	2.43E-07	5.12E-07	2.58	39.31	*TSG101, TFRC, SRC, CLTC, AGAP2, CXCR4, CBLB, AP2A1, SNX32, ARRB2, CBL, IL2RG, EGFR, PLD2, RAB11FIP1, GRK2, CXCR1, CXCR2, GRK6, PIP5K1B, CCR5, AP2M1, SMAD2, PDGFRA, RAB4A, SMAD3, SMURF1, CAV1, HSPA6, STAM, EPS15L1, TGFBR1, TGFBR2, DNM3, RAB31, ARPC2, BIN1, HGS, PARD3, TRAF6, IL2RB, MDM2, RAB5A, ARF5*

Thyroid hormone synthesis	20/75	6.25E-07	1.30E-06	4.40	62.90	*ATF2, GPX1, HSPA5, PRKCB, ADCY3, ITPR3, SERPINA7, ADCY7, TSHR, HSP90B1, CREB3, TPO, CREB1, CREB3L2, PRKACA, PLCB1, CGA, PLCB2, TSHB, ATF4*
Thermogenesis	40/232	1.18E-06	2.36E-06	2.54	34.64	*ATF2, PRKAA1, SMARCD3, KDM1A, UCP1, PRKAG2, ADCY3, ADCY7, NRAS, ATP5F1B, CPT2, RPS6KA2, CREB3L2, CYC1, PRKACA, PPARGC1A, CPT1A, ACSL1, ACTL6A, BMP8A, ACSL5, TSC2, ACSL4, GCG, ACSL3, SDHA, MAPK14, CPT1B, MAPK12, MTOR, SMARCA4, MAPK13, CREB3, MAPK11, CREB1, ADRB3, GRB2, PPARG, KRAS, NDUFAF1*
Oocyte meiosis	27/129	1.49E-06	2.97E-06	3.21	43.09	*CAMK2D, YWHAB, CUL1, ADCY3, ITPR3, ADCY7, CDC20, CCNB2, PPP2CB, CCNB1, RPS6KA2, MAPK1, PRKACA, BUB1, CAMK2G, SKP1, YWHAH, MAPK3, FBXW11, PPP2R5D, MAPK14, MAPK12, MAPK13, MOS, MAPK11, CCNE1, CPEB4*
Dopaminergic synapse	27/132	2.38E-06	4.68E-06	3.12	40.40	*ATF2, GSK3B, CAMK2D, ITPR3, ARRB2, GNAI2, MAPK9, PPP2CB, MAPK8, AKT2, CREB3L2, PRKACA, CAMK2G, PRKCB, PPP2R5D, FOS, PPP2R3A, MAPK14, MAPK12, MAPK13, MAPK10, CREB3, MAPK11, CREB1, PLCB1, PLCB2, ATF4*
Adrenergic signaling in cardiomyocytes	29/150	3.37E-06	6.59E-06	2.91	36.65	*ATF2, CAMK2D, CREM, ATP2A2, ADCY3, ADCY7, GNAI2, PPP2CB, RPS6KA5, AKT2, CREB3L2, MAPK1, PRKACA, CAMK2G, MAPK3, ATP2B4, PPP2R5D, PPP2R3A, MAPK14, AGT, MAPK12, MAPK13, CREB3, MAPK11, CREB1, BCL2, PLCB1, PLCB2, ATF4*
Aldosterone synthesis and secretion	22/98	4.10E-06	7.95E-06	3.51	43.49	*ATF1, ATF2, CAMK2D, PRKCB, ATP2B4, ADCY3, ITPR3, ADCY7, AGT, NR4A2, POMC, CREB3, CREB1, CREB3L2, ORAI1, PRKD1, PRKACA, PLCB1, PLCB2, CAMK2G, CAMK1G, ATF4*
Long-term potentiation	17/67	8.91E-06	1.72E-05	4.11	47.79	*CAMK2D, CREBBP, MAP2K2, PRKCB, ITPR3, NRAS, RPS6KA2, EP300, MAPK1, KRAS, PRKACA, PLCB1, RAF1, PLCB2, CAMK2G, ATF4, MAPK3*

Insulin secretion	19/86	2.32E-05	4.39E-05	3.43	36.60	*GLP1R, ATF2, SNAP25, CAMK2D, ABCC8, PRKCB, PDX1, ADCY3, ITPR3, GCG, ADCY7, CREB3, CREB1, CREB3L2, PRKACA, PLCB1, PLCB2, CAMK2G, ATF4*
Long-term depression	15/60	3.59E-05	6.67E-05	4.03	41.20	*LYN, GUCY1B1, MAP2K2, PRKCB, ITPR3, GNAI2, PPP2CB, NRAS, GNA12, MAPK1, KRAS, PLCB1, RAF1, PLCB2, MAPK3*
Renin secretion	16/69	5.40E-05	9.91E-05	3.65	35.83	*PTGER4, GUCY1B1, ACE, ITPR3, AGT, AQP1, GNAI2, EDNRA, CREB1, ADRB3, ADORA1, ORAI1, PRKACA, PLCB1, PLCB2, KCNJ2*
Cortisol synthesis and secretion	15/65	9.71E-05	1.77E-04	3.62	33.46	*ATF2, ADCY3, ITPR3, ADCY7, AGT, POMC, CREB3, CREB1, SP1, CREB3L2, ORAI1, PRKACA, PLCB1, PLCB2, ATF4*
Autoimmune thyroid disease	13/53	1.45E-04	2.58E-04	3.92	34.66	*CD86, IFNA5, IL10, IFNA16, CD80, PRF1, GZMB, TSHR, TPO, CD28, CTLA4, CGA, TSHB*
Viral myocarditis	14/60	1.45E-04	2.58E-04	3.67	32.46	*CD86, CD80, CAV1, ITGB2, PRF1, ITGAL, ICAM1, CASP8, CCND1, CASP3, RAC2, ABL1, CD28, FYN*
Ubiquitin mediated proteolysis	24/140	1.68E-04	2.95E-04	2.50	21.77	*PRKN, MAP3K1, UBA7, FBXW11, AIRE, SMURF1, CUL1, KEAP1, XIAP, CBLB, UBE2L6, CBL, PIAS2, PIAS1, CDC20, SOCS3, SOCS1, TRAF6, MDM2, BIRC6, STUB1, BIRC2, SKP1, BIRC3*
Circadian rhythm	9/31	3.95E-04	6.88E-04	4.93	38.61	*PRKAA1, CREB1, FBXW11, CUL1, RORC, PRKAG2, RORA, CSNK1D, SKP1*
Tight junction	26/169	5.36E-04	9.28E-04	2.20	16.57	*ITGB1, PRKAA1, ROCK1, ROCK2, SRC, PRKAG2, CD1D, F11R, CD1C, CD1B, MAPK9, PPP2CB, STK11, MAPK8, CCND1, ERBB2, PRKACA, JUN, MAP3K1, MSN, RUNX1, MAPK10, ARPC2, CDK4, PARD3, EZR*
Arrhythmogenic right ventricular cardiomyopathy	15/77	6.90E-04	1.18E-03	2.92	21.25	*ITGB1, TCF7L2, ITGA4, ITGA2, ITGA2B, LEF1, TCF7, ATP2A2, CDH2, ITGA11, ITGB8, CTNNB1, ITGB7, ITGA6, ITGA5*

Cocaine addiction	11/49	1.03E-03	1.75E-03	3.49	23.98	*ATF2, CREB3, JUN, CREB1, BDNF, CREB3L2, PRKACA, RELA, NFKB1, ATF4, GNAI2*
Type I diabetes mellitus	10/43	1.30E-03	2.19E-03	3.65	24.26	*CD86, IL1A, CD80, IL1B, ICA1, PRF1, CD28, IL12B, GZMB, IL12A*
Huntington disease	39/306	1.31E-03	2.19E-03	1.77	11.75	*HDAC1, CLTC, UCP1, HTT, AP2A1, MAPK9, MAPK8, ACTR1B, ATP5F1B, CASP8, POLR2A, ATG101, CASP3, CREB3L2, EP300, CYC1, PPARGC1A, AP2M1, DNAH12, CREBBP, GPX1, TBP, BDNF, TRAF2, SDHA, TUBB4A, MTOR, SOD1, MAPK10, CREB3, PSMA3, PSMC5, CREB1, SP1, BAX, PPARG, PLCB1, TP53, PLCB2*
Protein processing in endoplasmic reticulum	25/171	1.43E-03	2.37E-03	2.07	13.56	*ERO1A, PRKN, SAR1B, CUL1, HSP90B1, MAPK9, MAPK8, UFD1, CAPN1, UBQLN4, SKP1, TXNDC5, EDEM3, XBP1, HSP90AA1, HSPA5, HSPA6, TRAF2, MAPK10, BCL2, BAX, STUB1, ATF6, ATF4, NFE2L2*
Endocrine and other factor-regulated calcium reabsorption	11/53	2.04E-03	3.35E-03	3.15	19.55	*DNM3, PRKCB, VDR, CLTC, ATP2B4, AP2A1, PRKACA, PLCB1, PLCB2, ESR1, AP2M1*
Amphetamine addiction	13/69	2.10E-03	3.45E-03	2.80	17.25	*ATF2, CAMK2D, JUN, PRKCB, HDAC1, FOS, SIRT1, CREB3, CREB1, CREB3L2, PRKACA, CAMK2G, ATF4*
Circadian entrainment	16/97	2.91E-03	4.73E-03	2.38	13.91	*CAMK2D, GUCY1B1, PRKCB, ADCY3, ITPR3, FOS, ADCY7, GNAI2, RPS6KA5, CREB1, MAPK1, PRKACA, PLCB1, PLCB2, CAMK2G, MAPK3*
Hedgehog signalling pathway	11/56	3.22E-03	5.21E-03	2.94	16.90	*GSK3B, HHAT, GRK2, CCND1, FBXW11, SMURF1, CUL1, BCL2, CSNK1D, ARRB2, PRKACA*
Vibrio cholerae infection	10/50	4.26E-03	6.86E-03	3.01	16.43	*ERO1A, PLCG2, ADCY3, ATP6V1H, TCIRG1, PLCG1, PRKACA, ATP6V0C, ATP6V0A1, ATP6V1C2*
Parkinson disease	31/249	5.47E-03	8.78E-03	1.72	8.95	*PRKN, CAMK2D, UBA7, UBE2L6, ITPR3, PARK7, GNAI2, MAPK9, MAPK8, ATP5F1B, CASP3, PLCG1, CYC1, PRKACA, CAMK2G, SNCA, XBP1, HSPA5, DUSP1, SDHA, TUBB4A, MAPK10, PSMA3, PSMC5, ADORA2A, BAX, ATF6, TP53, ATF4, BCL2L1, NFE2L2*

Vasopressin-regulated water reabsorption	9/44	5.62E-03	8.96E-03	3.09	16.04	*CREB3, CREB1, CREB3L2, ADCY3, AQP4, STX4, AQP2, PRKACA, RAB5A*
Allograft rejection	8/38	7.38E-03	1.17E-02	3.21	15.75	*CD86, IL10, CD80, PRF1, CD28, IL12B, GZMB, IL12A*
Asthma	7/31	8.06E-03	1.27E-02	3.51	16.91	*IL10, CCL11, FCER1G, IL13, IL9, FCER1A, MS4A2*
Basal cell carcinoma	11/63	8.13E-03	1.27E-02	2.55	12.26	*GSK3B, TCF7L2, WNT10B, CDKN1A, APC, LEF1, TCF7, DVL2, BAX, CTNNB1, TP53*
Glutathione metabolism	10/57	1.10E-02	1.70E-02	2.56	11.55	*GSTM4, G6PD, GCLC, GPX1, RRM2, GSTO1, GSTP1, IDH1, IDH2, PRDX6*
Serotonergic synapse	16/113	1.28E-02	1.97E-02	1.99	8.66	*APP, PRKCB, DUSP1, ALOX15, ITPR3, PTGS2, GNAI2, NRAS, CASP3, MAPK1, KRAS, PRKACA, RAF1, PLCB1, PLCB2, MAPK3*
Gastric acid secretion	12/76	1.30E-02	1.98E-02	2.26	9.81	*CAMK2D, PRKCB, ADCY3, ITPR3, PRKACA, EZR, PLCB1, PLCB2, ADCY7, CAMK2G, KCNJ2, GNAI2*
Graft-versus-host disease	8/42	1.36E-02	2.07E-02	2.83	12.17	*CD86, IL1A, IL6, CD80, IL1B, PRF1, CD28, GZMB*
Amyotrophic lateral sclerosis	40/364	1.46E-02	2.21E-02	1.49	6.31	*PRKN, ITPR3, TANK, TBK1, ACTR1B, ATP5F1B, ATG101, CASP3, CASP1, CYC1, UBQLN4, DNAH12, XBP1, GPX1, HSPA5, NCBP1, NOS2, BAD, TRAF2, SDHA, TNFRSF1B, MAPK14, TUBB4A, MAPK12, MTOR, TNFRSF1A, MAPK13, SOD1, MAPK11, PSMA3, PSMC5, NRG3, BCL2, BAX, ATF6, TP53, SQSTM1, RAB5A, ATF4, BCL2L1*
Retrograde endocannabinoid signalling	19/148	1.92E-02	2.89E-02	1.78	7.02	*PRKCB, ADCY3, ABHD6, ITPR3, PTGS2, MAPK14, ADCY7, MAPK12, GNAI2, MAPK13, MAPK10, MAPK9, MAPK11, MAPK8, MAPK1, PRKACA, PLCB1, PLCB2, MAPK3*
Carbohydrate digestion and absorption	8/47	2.58E-02	3.87E-02	2.47	9.02	*PIK3CA, PRKCB, AKT2, PIK3CD, PIK3R2, PIK3R1, PLCB1, PLCB2*
Renin-angiotensin system	5/23	2.81E-02	4.20E-02	3.34	11.92	*CPA3, ACE, MME, LNPEP, AGT*

Glycine, serine and threonine metabolism	7/40	3.13E-02	4.65E-02	2.55	8.83	*GRHPR, ALAS2, ALAS1, GLDC, SHMT1, PHGDH, DLD*

Supplementary Table 4 Breakdown of machine learning tasks, methods, data types, results, cross validation (CV) usage, type of train test split, use of independent/external data and year published for each study that passed the inclusion/exclusion criteria of the systematic review.

AUC= Area under the curve, CV=cross-validation, IBD = Inflammatory Bowel Disease, CD = Crohn’s Disease, UC = Ulcerative Colitis, HC = Healthy Controls, GI = Gastrointestinal, LASSO = Least absolute shrinkage and selection operator, WGS = Whole Genome Sequencing, CAGI = Critical Assessment of Genome Interpretation

Paper	**Task**	**ML Type**	**ML Method(s)**	**Best ML Method**	**Study Size**	**Data Inc.**	**Type of Data**	**Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set**	**CV**	**Train/Test Split**	**External Test Data (Y/N)**	**Year**
Firouzi et al. [530]	Disease Course	Supervised	Decision Tree	Decision Tree	160 (121 UC, 39 CD)	IBD	Clinical	Accuracy=88.2% (UC), 89.8% (CD), 86.5% (IBD), Sensitivity=67.6% (UC), 82.8% (CD), 65.7% (IBD), Specificity=96.3% (UC), 95.2% (CD), 96.3% (IBD), Matthew's Correlation Coefficients=0.69 (UC), 0.79 (CD), 0.68 (IBD)	10-fold CV	No	N	2007

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
Waljee et al.[268]	Treatment Response	Supervised	Random Forest	Random Forest	346	IBD	Clinical	AUC=0.856 (response), 0.813 (non-adherence), 0.797 (shunting)	10-fold CV on training data	Yes (hold out validation)	Y	2009
Cui et al.[513]	Diagnosis	Supervised	Support Vector Machine	Support Vector Machine	124 (25 IBD, 99 HC)	IBD	Metagenomic	Accuracy=88%, Sensitivity=92%, Specificity=84%	Leave-one-out CV	Yes (hold out validation)	N	2013
Wei et al.[436]	Risk of Disease	Supervised	Logistic Regression	Logistic Regression	53,279 (17,379 CD, 13,458 UC, 22,442 HC)	IBD	Genome wide	Cross Validation: AUC=0.864 (CD) 0.83 (UC). Independent Test Set: AUC=0.864 (CD), 0.826 (UC)	10-fold CV	Yes (hold out validation)	N	2013
Hübenthal et al.[512]	Diagnosis	Supervised	Support Vector Machine, Random Forest	Random Forest	273 (37 CD, 32 UC, 92 HC, 113 other)	IBD	MicroRNA Expression	AUC=0.996	5-fold CV on training data	Yes (hold out validation)	N	2015
Niehaus et al.[501]	Disease Severity	Supervised and Unsupervised	Logistic Regression, Random Forests, Support Vector Machine, Hierarchical Clustering	Support Vector Machine, Hierarchical Clustering	501	CD	Health Records, EMR Databases	Accuracy=68.7%, Sensitivity=59.1%, Specificity=78.4%	5-fold CV on training data	Yes (hold out validation)	N	2015

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
Menti et al.[434]	Disease Severity - Complications	Supervised	Naïve Bayes, Bayesian Additive Regression Trees, Bayesian Networks	Bayesian Networks	152	IBD	Clinical, Genetic	AUC=0.95, Accuracy=0.89, Sensitivity=0.78, Specificity=0.94	10-fold CV	No	N	2016
Eck et al.[262]	Diagnosis	Supervised	Support Vector Machine, Random Forest, Nearest Shrunken Centroids, Logistic Regression	Support Vector Machine, Random Forest	112 (56 IBD, 56 HC)	IBD	Microbiota	Accuracy=81%	10-fold CV	No	N	2017
Waljee et al.[528]	Disease Course	Supervised	Logistic Regression, Random Forest	Random Forest	20368	IBD	Clinical	Predict Hospitalisation and Corticosteroid Prescriptions. IBD: AUC=0.87, Sensitivity=74-80%, Specificity=80-82%. UC: AUC=0.84. CD=0.85. IC=0·82. Predict Corticosteroid Prescription Only, IBD: AUC=0.9. Predict Hospitalisation and Corticosteroid Prescriptions (12 month outcome): AUC=0.9	None	Yes (hold out validation)	N	2017

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
Yu et al.[554]	Identification of Patients	Supervised	Natural Language Processing	Natural Language Processing	2393 (600 UC, 600 CD, 1193 Other)	IBD	Electronic Medical Records	CD AUC~0.95, UC AUC~0.95, CD F-score~0.83, UC F-score~0.89	None	Out of sample accuracy	N	2017
Isakov et al.[553]	Risk of Disease	Supervised	Random Forest, Support Vector Machine, Extreme Gradient Boosting, Elastic net regularised generalised linear model, Combined Model (elastic net regularised generalised linear model, extreme gradient boosting, SVM, RF)	Combined Model (elastic net regularised generalised linear model, extreme gradient boosting, SVM, RF)	513 (180 CD, 149 UC, 94 colorectal neoplasms, 90 normal tissue)	IBD	Gene Expression	AUC=0.829, Accuracy=0.808, Sensitivity=0.577, Specificity=0.880	5-fold CV	Yes (hold out validation)	N	2017
Pal et al.[545]	Risk of Disease	Supervised	Naïve Bayes, Logistic Regression, Consensus Method (Naïve Bayes, Logistic Regression, Random Forest)	Consensus Method (Naïve Bayes, Logistic Regression, Random Forest)	111 (64 CD, 47 HC)	CD	Genome wide association study, exome sequencing to	AUC=0.72	None	Yes (hold out validation)	N	2017

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
							impute genotypes.					
Daneshjou et al.[547]	Risk of Disease	Supervised	Metaclassifier	Metaclassifier	111 (64 CD, 47 HC)	CD	Exome sequencing data	AUC=0.78	CV performed	Splits performed	N	2017
Giollo et al.[439]	Risk of Disease	Supervised	Support Vector Machine, Ensemble Classifier	Ensemble Classifier	111 (64 cases, 47 controls)	CD	Genetic	AUC=0.66	None	Yes (hold out validation)	N	2017
Mossotto et al.[254]	Subtype Diagnosis	Supervised and Unsupervised	Support Vector Machine, Hierarchical Clustering	Support Vector Machine, Hierarchical Clustering	287 (178 CD, 80 UC, 29 IBDU (only reclassified))	IBD	Clinical	Cross-validation: AUC=0.87, Accuracy=82.7%, Precision=0.91, Recall=0.83, F1-score=0.87. Test set: Accuracy=83.3%, Precision=0.86, Recall=0.83, F1-score=0.84	5-fold cross validation on training data	Yes (hold out validation)	N	2017
Waljee et al.[552]	Treatment Response	Supervised	Random Forest	Random Forest	1080	IBD	Clinical	Objective Remission: AUC=0.79, Sensitivity=70.6%, Specificity=73.8%. Non-adherence: AUC=0.84, Sensitivity=70.6%, Specificity=85.0%. Shunting:	Out of bag error	Yes (hold out validation)	N	2017

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
								AUC=0.78, Sensitivity=65.2%, Specificity=79.0%.				
Kang et al.[260]	Treatment Response	Supervised	Gene Regulatory Network-based Regularized Artificial Neural Network (GRRANN), Group LASSO, Regularized Logistic Regression, Multilayer Perceptron, Support Vector Machine	Gene Regulatory Network-based Regularized Artificial Neural Network (GRRANN)	46	UC	Gene Expression	Balanced Accuracy≈0.8	5-fold CV on training data	Yes (hold out validation)	N	2017
Forbes et al.[272]	Diagnosis	Supervised	Random Forest	Random Forest	102 (20 CD, 19 UC, 23 HC, 40 Other)	IBD	Microbiota	Diseased vs HC: AUC=0.93, Balanced Accuracy=0.84. Breakdown per inflammatory disease found in paper.	Out of bag error	No	N	2018
Douglas et al.[511]	Diagnosis, Treatment Response	Supervised	Random Forest	Random Forest	771. Validation data (diagnosis only): 731 (444	CD	Metagenomic	Diagnosis: Accuracy=84.2%. Independent Validation (diagnosis): Accuracy 73.2%.	Out of bag error, Leave-	No	Y	2018

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
					CD, 287 control)			Treatment Response: Accuracy 77.8%.	one-out cross-validation			
Jain et al.[529]	Disease Course	Supervised	Random Forest	Random Forest	179	UC	Clinical	Colectomy Prediction: Accuracy=77%, Sensitivity=75%, Specificity=80%. Steroid Dependence: Accuracy=75%, Sensitivity=69%, Specificity=80%.	None	Yes (hold out validation)	N	2018
Reddy et al.[574]	Disease Severity - Activity	Supervised	Gradient Boost, Logistic Regression, Regularised Regression	Gradient Boost	82 CD	CD	Clinical (EHR)	AUC=0.93 to predict disease severity using C-reactive protein as proxy	10-fold CV replicated 10 times	No	N	2018
Maeda et al.[575]	Disease Severity - Activity	Supervised	Support Vector Machine	Support Vector Machine	187 patients, 22835 images	UC	Endocytoscopic Image	Accuracy=91%, Kappa=1, Sensitivity=74%, Specificity=97%	None	Yes (hold out validation)	N	2018
Han et al.[274]	Subtype Diagnosis	Supervised	Random Forest	Random Forest	163 (Train: 24 CD, 59 UC, Validation set 1:5 CD ,7 UC.	IBD	Gene Expression	Validation set 1: AUC=0.829. Validation set 2: AUC=0.764. Validation set 3: AUC=0.836. Validation set 4: AUC=0.849	None	Yes (train on one dataset,test	Y	2018

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
					Validation set 2: 14 CD, 10 UC. Validation set 3: 11 CD, 5 UC. Validation set 4: 13 CD, 15 UC)					on four externals)		
Waljee et al.[548]	Treatment Response	Supervised	Random Forest	Random Forest	491	UC	Clinical	AUC=0.73, Sensitivity=0.72, Specificity=0.68	None	Yes (hold out validation)	N	2018
Doherty et al.[546]	Treatment Response	Supervised	Random Forest	Random Forest	306 (232 treated, 74 untreated)	CD	Microbial Genome, Clinical	Remission: AUC=0.844, Sensitivity=0.774, Specificity=0.831. Response: AUC=0.733 Sensitivity=0.684, Specificity=0.724	None	No	N	2018
Romagnoni et al.[437]	Diagnosis	Supervised	Logistic Regression, Gradient Boosted Trees, Artificial Neural Network	Logistic Regression	52277 (18,227 CD, 34,050 HC)	CD	Genomic (Immunochip)	AUC=0.8	10-fold CV on training data	Yes (hold out validation)	N	2019

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
Wang et al.[440]	Diagnosis	Supervised	Analysis of Variation for Association with Disease (Support Vector Machine Based)	Analysis of Variation for Association with Disease (Support Vector Machine Based)	173 (111 CD, 62 HC)	CD	Genomic (WES)	AUC=0.75	Leave-one-out CV on training data	Yes (CV on one data set, test on external)	Y	2019
Morell Miranda et al.[527]	Disease Course	Supervised	Random Forest	Random Forest	70 patients, 1084 metagenomic samples, 566 metatranscriptomic samples	IBD	Metagenomic, Metatranscriptomic	Micro-averaged AUC=0.96 (metagenomic data), AUC=0.91 (metatranscriptomic data), AUC=0.99 (combined data)	1000-fold CV (combined data), 500-fold CV (individual datasets)	Yes (hold out validation)	N	2019
Bottigliengo et al.[435]	Disease Course - Extra Intestinal Manifestations	Supervised	Naïve Bayes, Bayes Network, Bayes additive regression trees	Naïve Bayes, Bayes Network, Bayes additive regression trees	152	CD	Clinical, SNP Panel	AUC=0.75	10-fold CV, replicated 10 times	No, used 1000 bootstrapped samples	N	2019
Braun et al.[525]	Disease Course - Relapse	Supervised	Random Forest	Random Forest	45 patients, 217 samples	CD	Metagenomic, Clinical	Relapsers vs Non-relapsers: AUC=0.78.	Out of Bag Error	No	N	2019

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
Waljee et al.[432]	Disease Course - Remission	Supervised	Random Forest	Random Forest	401 IBD	IBD	Clinical	C-reactive protein as marker of remission, AUC=0.78	None	Yes, hold out replicated 100 times	N	2019
Dong et al.[433]	Disease Course - Surgery	Supervised	Random Forest, Logistic Regression, Support Vector Machine, Decision Tree, Artificial Neural Network	Random Forest	239 CD (83 surgery, 156 no surgery)	CD	Clinical	Accuracy=0.96, F1=0.77, AUC=0.98	10-fold CV on train and test data	Yes (hold out validation)	N	2019
Biasci et al.[538]	Disease Severity	Supervised	Elastic Net	Elastic Net	118 (66 CD, 52 UC). External Data: 123 (66 CD 57 UC)	IBD	Expression (qPCR)	Accuracy=0.81	Leave-one-out CV	Yes (CV on one data set, test on external)	Y	2019
Lerrigo et al.[497]	Patient clustering	Unsupervised	Latent Dirichlet Allocation	Latent Dirichlet Allocation	51,591 entries	IBD	Online Posts	Identified most common emotions with IBD patients.	NA	NA	N	2019
Taylor et al.[438]	Risk of Disease	Supervised	Elastic net, Random Forest	Random Forest	454 CD (124 used)	CD	Clinical, Genotyping	AUC=0.87	20 repeats of 5-fold CV on training data	Yes (hold out validation)	N	2019

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
Khorasani et al.[507]	Diagnosis	Supervised	Support Vector Machine	Support Vector Machine	77 in training data (39 UC, 38 controls), 85 in active test data (74 UC, 11 controls), 34 in inactive test data (23 UC, 11 controls)	UC	Gene Expression	Inactive UC vs Controls : Precision-recall AUC=0.68. Active UC vs Controls : Precision-recall AUC 1.	5-fold CV on training data	Yes (hold out validation)	N	2020
Raimondi et al.[441]	Diagnosis	Supervised	Neural Network	Neural Network	CAGI 2 (42 cases, 14 controls), CAGI 3 (51 cases, 15 controls), CAGI 4 (64 cases, 47 controls)	CD	Whole Exome Sequencing	AUC=82.5, Sensitivity=96.2, Specificity=60.0, Precision=89.3, Precision-recall AUC=93.1.	Leave-one-out CV	Yes (hold out validation)	N	2020
Jiang et al.[508]	Diagnosis	Supervised	Random Forest	Random Forest	492 (110 CD, 382 HC)	CD	Metagenomic	AUC=0.92-0.95	10-fold CV (leave one dataset out -	No	N	2020

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
									pooled many datasets)			
Iablokov et al.[510]	Diagnosis	Supervised	Random Forest	Random Forest	654 HC, 274 CD, 175 UC	IBD	Metagenomic	HC vs CD: taxonomy-based classifier AUC=0.79-0.95, phenotype-based classifier AUC=0.76-0.91. HC vs UC: taxonomy-based AUC=0.65-0.92, phenotype-based AUC=0.43-0.92	3-fold leave one dataset out	Yes (hold out validation)	N	2020
Clooney et al.[498]	Diagnosis, Disease Course - Remission	Supervised and Unsupervised	Gradient boosted trees, hierarchical clustering	Gradient boosted trees, hierarchical clustering	692 (303 CD, 228 UC, 161 controls)	IBD	16S rRNA (from faecal microbiota)	CD vs HC: AUC=0.88, Accuracy=84%. UC vs HC AUC=0.88, Accuracy=83%. CD vs UC AUC=0.67, Accuracy=64%. Inactive vs Active CD AUC=0.81, Accuracy=81%. Inactive vs Active UC AUC=0.73, Accuracy=85%. Inactive vs	Parameter optimisation: bootstrapping with 1000 iterations and 5-fold CV. Classification: leave-one-out CV	No	N	2020

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
								Active IBD AUC=0.91, Accuracy=89%. Clustering revealed 10 subgroups.				
Shivaji et al.[520]	Disease Course	Supervised	Logistic Regression	Logistic Regression	147	IBD	Clinical, Demographic	Used to identify variables which are significantly predictive of outcomes after biologic therapy was discontinued.	CV according to CARRoT package (R)	No	N	2020
Waljee et al.[514]	Disease Course - Remission	Supervised	Random Forest, LASSO Logistic Regression	LASSO Logistic Regression	117	CD	Clinical	Clinical Remission (CD activity index): AUC=0.61. Endoscopic Remission (Faecal calprotectin) AUC=0.6. Biological Remission (C-reactive protein) AUC=0.62.	CV on training data	Yes (hold out validation)	N	2020
Sakurai et al.[517]	Disease Course - Remission	Supervised	Logistic Regression, Naïve Bayes, Neural Network, Support Vector Machine, AdaBoost, CN2 rule inducer, Tree, Random	Logistic Regression, Naïve Bayes, Neural Network, Support Vector Machine	12 (9 UC, 3 normal)	UC	Gene Expression	For all ML methods listed in Best ML method, for normal vs relapse at week 0 vs non-relapse at week 0: AUC=1, F1=1, Precision=1, Recall=1.	10-fold CV	No	N	2020

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
			Forest, k-Nearest Neighbours									
Taylor et al.[521]	Disease Course - Remission	Supervised	Partial Least Squares Discriminant Analysis	Partial Least Squares Discriminant Analysis	25 patients (11 low calprotectin, 11 high calprotectin)	CD	NMR Spectra	R^2=0.87 (goodness of fit on training data), Q^2=0.41 (goodness of prediction on test data).	Monte Carlo CV	Yes (hold out validation)	N	2020
Jones et al.[524]	Disease Course - Remission	Supervised	Random Forest	Random Forest	18 patients, 139 samples	CD	16S Microbiome, Clinical	AUC=0.9	Leave-one-out CV	No	N	2020
Takenaka et al.[526]	Disease Course - Remission	Supervised	Deep Neural Net	Deep Neural Net	40,758 images and 6885 biopsies for training, 4187 images and 4104 biopsies for testing	UC	Endoscopy Images, Histology	Prediction of remission by endoscopic and histological state looking at endoscopy images. Endoscopic remission: Accuracy=0.9, k coefficient=0.8, Histology remission: Accuracy=0.93, k coefficient=0.86	None	Yes (hold out validation)	N	2020

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
Choi et al.[522]	Disease Course - Required Treatment	Supervised	Logistic Regression, Support Vector Machine, Random Forest, XGBoost, artificial neural network	Ensemble (XGB and ANN are base learners here)	Dataset 1 (GMC): 1299 patients (135 biologic). External Data (K-CDM): 1987 (146 biologic)	IBD	Clinical	Test data AUC=0.86. K-CDM dataset external validation AUC=0.81.	5-fold on training data	Yes (hold out validation)	Y	2020
Ghoshal et al.[523]	Disease Course - Surgery	Supervised	Artificial Neural Network	Artificial Neural Network	263 (231 responders, 28 non-responders)	UC	Clinical	Accuracy=73% in classifying response to medical treatment	None	Yes (hold out validation)	N	2020
Sofo et al.[519]	Disease Course - Surgery Complications	Supervised	Support Vector Machine	Support Vector Machine	32	UC	Demographic, Clinical	Predicting infectious minor complications: Strike rate=84.3%, Sensitivity=87.5%, Specificity=83.3%	Leave-one-out CV on training data, no separate test data	No	N	2020
Popa et al.[537]	Disease Severity - Activity	Supervised	Neural Network	Neural Network	55 UC	UC	Clinical, Endoscopy	Classify active disease at 1 year. AUC=0.92 on test and AUC=1 on validation (5 samples).	10-fold CV on training data	Yes (hold out validation)	N	2020

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
Bossuyt et al.[541]	Disease Severity - Activity	Supervised	Multiple Regression	Multiple Regression	29 UC, 6 HC	UC	Endoscopy Images	UC vs healthy tissue using the red channel.	None	No	N	2020
Yao et al.[534]	Disease Severity - Activity	Supervised	Convolutional Neural Network	Convolutional Neural Network	51 videos, 34,810 frames (training data). 124 videos (test data)	UC	Endoscopy Video	Mayo endoscopy score prediction. 5-fold CV training: Accuracy=0.876, Sensitivity=0.902, Specificity=0.87, AUC=0.961, F1=0.834, Precision=0.79, Average precision=0.932, Independent test data: Accuracy=0.844, Sensitivity=0.834, Specificity=0.851, AUC=0.93, F1=0.804, Precision=0.831, Average precision=0.91	5-fold CV	Yes (CV on one data set, test on external)	Y	2020
Gottlieb et al.[535]	Disease Severity - Activity	Supervised	Recurrent Neural Network	Recurrent Neural Network	786 videos, 7,400,000 frames	UC	Endoscopy Video	Quadratic weighted kappa (QWK)=0.844 for the outcome endoscopic mayo score,	5-fold CV, hold-out	Yes (hold out validation)	N	2020

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
								QWK=0.855 for the outcome UC endoscopic index of severity.				
Wang et al.[536]	Disease Severity - Complications	Supervised	Decision Tree	Decision Tree	67 CD	CD	Clinical	Stricturing vs not-stricturing: AUC=0.917 to	None	Yes (hold out validation)	N	2020
Ungaro et al.[485]	Disease Severity - Complications	Supervised	Random Survival Forest	Random Survival Forest	265 (98 with complications, 167 no complications)	CD	Protein Biomarkers, Clinical	Any complication AUC=0.69.B2 complications (stricturing) AUC=0.70. B3 complications (penetrating) AUC=0.79.	Out of bag performance measures, 5-fold CV, 200 replication	No	N	2020
Wang et al.[544]	Medication Adherence	Supervised	Back-propagation neural network, Support Vector Machine, Logistic Regression	Support Vector Machine	446 CD	CD	Clinical	Accuracy=87.7%	10-fold CV	No	N	2020
Kieft et al.[496]	Patient clustering	Unsupervised	Support Vector Machine, Random Forest, Neural Network	Neural Network	102 (49 CD, 53 HC). External Data: 64 (43 CD, 21 HC)	CD	Metagenomic	Identified differentially abundant virus and then clustered them. Some classes of virus more abundant in CD	NA	NA	Y	2020

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
								(Fusobacteria-like and Enterobacterales).				
Coelho et al.[391]	Patient clustering	Unsupervised	Hierarchical Clustering	Hierarchical Clustering	22	IBD	Immunoassay	Groups identified do not correlate with clinical phenotypes	NA	NA	N	2020
Le et al.[500]	Predict metabolite abundance from microbiome	Supervised and Unsupervised	Sparse Neural Encoder-Decoder Network	Sparse Neural Encoder-Decoder Network	.	CD	Microbiome, Metabolomic	Use microbiome to predict metabolites abundance, then use to cluster and predict: CD vs HC ~0.94.	5-fold CV	No	N	2020
Tong et al.[549]	Subtype Diagnosis	Supervised	Random Forest, Convolutional Neural Network	Random Forest	6399 (5128 UC, 875 CD, 396 ITB)	IBD	Clinical	UC vs CD Sensitivity=0.89, Specificity=0.84, AUC=0.94	10-fold CV on training data	Yes (hold out validation)	N	2020
McDonnell et al.[550]	Treatment Response	Supervised	Random Forest Regression	Random Forest Regression	94 (54 CD, 36 UC, 4 IBDU)	IBD	Clinical	Mean squared error=1.876.	5-fold CV on testing data	Yes (hold out validation)	N	2020
Biernacka et al.[502]	Diagnosis	Supervised	Kohonene Neural Network	Kohonene Neural Network	131 (60 CD, 17 UC, 26	IBD	Imaging (Magnetic	Identified features associated with CD (intended for triaging	None	No	N	2021

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
					suspected, 28 other GI/unknown)		Resonance Enterography)	unwell patients to more invasive assessment).				
Volkova et al.[503]	Diagnosis	Supervised	Random Forest, eXtreme Gradient Boosting (XGBoost), Ridge Regression, Support Vector Machine (radial kernel)	XGBoost	894 16S rRNA, 578 Metagenomics	IBD	16S rRNA Sequencing, Shotgun Metagenomics	IBD vs HC 16S (species): AUC=0.942, F1=0.682. IBD vs HC metagenomics (species): AUC=0.950, F1=0.680.	7-fold-3-times CV on training data	Yes (hold out validation)	N	2021
Sarrabayrouse et al.[509]	Diagnosis, Disease Course - Relapse	Supervised	Random Forest	Random Forest	206 (86 HC, 89 CD, 31 UC)	IBD	Metagenomic, Clinical	IBD vs Control AUC=0.84, also individual models generated, but less performant.	None	Yes (hold out validation)	N	2021
Nuzzo et al.[504]	Diagnosis, Subtype Diagnosis	Supervised	Logistic regression, k-Nearest Neighbours, random forest, 3-layer neural net, naïve Bayes, linear kernel one-vs-rest, XGBoost, generalised mixed effects random forest	XGBoost	252 (127 CD, 74 UC, 51 non-IBD) samples	IBD	Metabolomics	Method performance assessed with F1 score	10-fold Stratified CV on training set (XGBoost had additional 5-fold CV for hyperparameter tuning)	Yes (hold out validation)	N	2021

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
Xu et al.[505]	Diagnosis, Subtype Diagnosis	Supervised	Logistic regression, random forest, gradient boosting classifier, support vector machine, LightGBM	LightCUD (based on LightGBM)	349 samples, 271 individuals (127 UC, 21 CD, 201 HC). Independent data: 185 (16 HC, 169 CD)	IBD	16S and Whole Genome Microbial Sequencing	5-fold CV training data results (WGS-based modules). IBD vs HC: AUC=0.984, Average precision=0.947. CD vs UC: AUC=0.966, Average precision=0.953. Test Set results: CD vs HC AUC=0.809, Average precision=0.971; CD vs UC accuracy=76.9%	5-fold CV	Yes (CV on one data set, test on external)	Y	2021
Manandhar et al.[506]	Diagnosis, Subtype Diagnosis	Supervised	Random forest, Decision Tree, Elastic Net, Support Vector Machine (radial), Neural networks	Random Forest	IBD vs non-IBD: 1429 (729 IBD, 700 non-IBD). CD vs UC: 585 (406 CD, 179 UC)	IBD	Faecal 16S Metagenomic	IBD vs non-IBD: AUC=0.82, Accuracy=0.74, Sensitivity=0.84, Specificity=0.64, Precision=0.7, F1=0.76. CD vs UC: AUC=0.92, Accuracy=0.83, Sensitivity=0.85, Specificity=0.80, Precision=0.9, F1=0.88.	10-fold CV on training data	Yes (hold out validation)	N	2021

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
Udristoiu et al.[516]	Disease Course - Remission	Supervised	Convolutional Neural Network, Convolutional Neural Networks with Long Short-Term Memory (Recurrent Neural Network)	Convolutional Neural Networks with Long Short-Term Memory	54 patients (active CD 32, controls 22 (18 CD w/ mucosal healing, 4 no IBD normal mucosa). 6205 images (3672 active, 2533 no inflammation)	CD	Endomicroscopy Images	Normal vs inflamed colonic mucosa: Accuracy=95.3%, Specificity=92.78%, Sensitivity=94.6%, AUC=0.98, Precision-Recall AUC=0.93.	None	Yes (hold out validation)	N	2021
Stidham et al.[515]	Disease Course - Surgery	Supervised	LASSO Logistic Regression, Random Forest	LASSO Logistic Regression	2809 patients, 4950 observations (256 Surgery)	CD	Demographic, Clinical	Average AUC=0.78, Sensitivity=0.735, Specificity=0.726.	5-fold CV on training data, 10-fold CV on test data	Yes (hold out validation)	N	2021
Kang et al.[518]	Disease Course - Surgery	Supervised	CatBoost (Tree Based)	CatBoost (Tree Based)	337 (46 with intestinal resection, 291 controls). External data: 126 (19	CD	Clinical, SNP Genotype	Predicting early intestinal resection. Internal validation: AUC=0.878, External validation: AUC=0.836	5-fold CV on training data	Yes (hold out validation)	Y	2021

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
					intestinal resection)							
Dorofeyev et al.[531]	Disease Severity	Supervised	Intelligent Monitoring	Intelligent Monitoring	223 (104 CD, 79 UC, 20 HC, 20 non-IBD controls)	IBD	Clinical	Correctly classified 92% of points.	NA	NA	N	2021
Mohapatra et al.[539]	Disease Severity	Supervised	DeepCNN	DeepCNN	1000 images	IBD	Endoscopy Images	Accuracy=0.94, F1=0.94, Precision=0.94, Recall=0.94, Specificity=0.99	None	Yes (hold out validation)	N	2021
Gutierrez Becker et al.[532]	Disease Severity - Activity	Supervised	Convolutional Neural Network	Convolutional Neural Network	4371 training frames, 1672 videos, 1105 patients (test data 1). 778 still frames (test data 2)	UC	Endoscopy Video	Mayo clinic endoscopic subscore (MCES) prediction on 5-fold CV training data (each are binary classification tasks): MCES ⩾ 1 AUC=0.84, Precision=0.92, Recall=0.79; MCES ⩾ 2 AUC=0.85, Precision=0.85, Recall=0.81; MCES ⩾ 3 AUC=0.85,	5-fold CV	Yes (CV on one data set, test on external)	Y	2021

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
								Precision=0.81, Recall=0.77. External test set: MCES ⩾ 2 AUC=0.82, Precision=0.92, Recall=0.73; MCES ⩾ 3 AUC=0.83, Precision=0.39, Recall=0.84.				
Takenaka et al.[540]	Disease Severity - Activity	Supervised	Deep Neural Network (DNUC)	Deep Neural Network (DNUC)	875 patients	UC	Endoscopy Images	Evaluating mucosal healing: Sensitivity=92.0%, Specificity 91.3%, Positive predictive value=86.2%, Negative predictive value=95.1%	NA	NA	NA	2021
Li et al.[533]	Disease Severity - Complications	Supervised	Radiomic model (Logistic regression)	Radiomic model	167 patients, 212 lesions	CD	Computed-Tomography Enterography Imaging	Moderate-severe vs none-mild intestinal fibrosis. Test cohort performance in 3 referral centres AUC=0.816, AUC=0.724, AUC=0.750.	Leave-one-out CV on training data	Yes (multiple hold out sets)	N	2021
Liu et al.[495]	Patient clustering	Unsupervised	Guassian Mixture Model	Guassian Mixture Model	1961 (843 UC, 1118 CD)	IBD	Questionnaire	Identified two clusters, performed genome-wide association study on these groups.	NA	NA	N	2021

Paper	Task	ML Type	ML Method(s)	Best ML Method	Study Size	Data Inc.	Type of Data	Best Results (Metrics) Reported from validation or cross-validation, and where conducted, the test set	CV	Train/Test Split	External Test Data (Y/N)	Year
Dhaliwal et al.[499]	Subtype Diagnosis	Supervised and Unsupervised	Random Forest, Similarity Network Fusion Clustering	Random Forest, Similarity Network Fusion Clustering	58 colonic IBD (41 UC, 17 CD) in training data, 15 IBD in testing data	IBD	Clinical	Accuracy=97% in CV training. Accuracy=100% in hold-out test set. Unsupervised clustering identified two groups: group 1 55 patients (98% UC), group 2 18 patients (94% colonic CD). Two samples misclassified (as in RF modelling)	Leave-one-out CV	Yes (hold out validation)	N	2021
Jiang et al.[551]	Subtype Diagnosis	Supervised	Random Forest	Random Forest	763 Cases (CD, UC, Colorectal Cancer), 632 HC	IBD	Metagenomic	Multivariate analysis to identify a set of metagenomic markers, then use markers to build multiclass: AUC=0.75-0.9; Case-control classifier AUC=0.88.	10-fold CV replicated 10 times and leave one dataset out for validation	No	N	2021

List of References

[1] Bonen DK, Cho JH. The genetics of inflammatory bowel disease. Gastroenterology. 2003;124(2):521-36.
[2] Ogura Y, Bonen DK, Inohara N, Nicolae DL, Chen FF, Ramos R, et al. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. Nature. 2001;411(6837):603-6.
[3] Hugot J-P, Chamaillard M, Zouali H, Lesage S, Cézard J-P, Belaiche J, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. Nature. 2001;411(6837):599-603.
[4] Ramos GP, Papadakis KA. Mechanisms of Disease: Inflammatory Bowel Diseases. Mayo Clinic proceedings. 2019;94(1):155-65.
[5] Ashton JJ, Mossotto E, Stafford IS, Haggarty R, Coelho TAF, Batra A, et al. Genetic Sequencing of Pediatric Patients Identifies Mutations in Monogenic Inflammatory Bowel Disease Genes that Translate to Distinct Clinical Phenotypes. Clin Transl Gastroenterol. 2020;11(2):e00129-e.
[6] Worthey EA, Mayer AN, Syverson GD, Helbling D, Bonacci BB, Decker B, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. Genet Med. 2011;13(3):255-62.
[7] Levine A, Koletzko S, Turner D, Escher JC, Cucchiara S, de Ridder L, et al. ESPGHAN Revised Porto Criteria for the Diagnosis of Inflammatory Bowel Disease in Children and Adolescents. Journal of Pediatric Gastroenterology and Nutrition. 2014;58(6).
[8] Lamb CA, Kennedy NA, Raine T, Hendy PA, Smith PJ, Limdi JK, et al. British Society of Gastroenterology consensus guidelines on the management of inflammatory bowel disease in adults. Gut. 2019;68(Suppl 3):s1.
[9] Diefenbach KA, Breuer CK. Pediatric inflammatory bowel disease. World journal of gastroenterology. 2006;12(20):3204-12.
[10] Dahlhamer JM ZE, Ward BW, Wheaton AG, Croft JB. Prevalence of Inflammatory Bowel Disease Among Adults Aged ≥18 Years — United States, 2015. MMWR Morb Mortal Wkly Rep. 2016;65:1166 – 9.
[11] Jones G-R, Lyons M, Plevris N, Jenkinson PW, Bisset C, Burgess C, et al. IBD prevalence in Lothian, Scotland, derived by capture–recapture methodology. Gut. 2019;68(11):1953.
[12] Kaplan GG, Bernstein CN, Coward S, Bitton A, Murthy SK, Nguyen GC, et al. The Impact of Inflammatory Bowel Disease in Canada 2018: Epidemiology. Journal of the Canadian Association of Gastroenterology. 2018;2(Supplement_1):S6-S16.
[13] Ashton JJ, Wiskin AE, Ennis S, Batra A, Afzal NA, Beattie RM. Rising incidence of paediatric inflammatory bowel disease (PIBD) in Wessex, Southern England. Archives of Disease in Childhood. 2014;99(7):659.
[14] Ashton JJ, Cullen M, Afzal NA, Coelho T, Batra A, Beattie RM. Is the incidence of paediatric inflammatory bowel disease still increasing? Archives of Disease in Childhood. 2018;103(11):1093.
[15] Ashton JJ, Barakat FM, Barnes C, Coelho TAF, Batra A, Afzal NA, et al. Incidence and Prevalence of Paediatric Inflammatory Bowel Disease Continues to Increase in the South of England. Journal of Pediatric Gastroenterology and Nutrition. 2022;75(2).
[16] Pasvol TJ, Horsfall L, Bloom S, Segal AW, Sabin C, Field N, et al. Incidence and prevalence of inflammatory bowel disease in UK primary care: a population-based cohort study. BMJ Open. 2020;10(7):e036584.
[17] Mak WY, Zhao M, Ng SC, Burisch J. The epidemiology of inflammatory bowel disease: East meets west. Journal of Gastroenterology and Hepatology. 2020;35(3):380-9.
[18] Baumgart DC, Sandborn WJ. Inflammatory bowel disease: clinical aspects and established and evolving therapies. Lancet (London, England). 2007;369(9573):1641-57.

[19] Van Limbergen J, Russell RK, Drummond HE, Aldhous MC, Round NK, Nimmo ER, et al. Definition of phenotypic characteristics of childhood-onset inflammatory bowel disease. Gastroenterology. 2008;135(4):1114-22.
[20] Yu YR, Rodriguez JR. Clinical presentation of Crohn’s, ulcerative colitis, and indeterminate colitis: Symptoms, extraintestinal manifestations, and disease phenotypes. Seminars in Pediatric Surgery. 2017;26(6):349-55.
[21] Ricciuto A, Fish JR, Tomalty DE, Carman N, Crowley E, Popalis C, et al. Diagnostic delay in Canadian children with inflammatory bowel disease is more common in Crohn's disease and associated with decreased height. Arch Dis Child. 2018;103(4):319-26.
[22] Gajendran M, Loganathan P, Catinella AP, Hashash JG. A comprehensive review and update on Crohn's disease. Disease-a-Month. 2018;64(2):20-57.
[23] Magro F, Langner C, Driessen A, Ensari A, Geboes K, Mantzaris GJ, et al. European consensus on the histopathology of inflammatory bowel disease☆. Journal of Crohn's and Colitis. 2013;7(10):827-51.
[24] Feakins RM. Inflammatory Bowel Disease Diagnosis. In: Feakins RM, editor. Non-Neoplastic Pathology of the Gastrointestinal Tract: A Practical Guide to Biopsy Diagnosis. Cambridge: Cambridge University Press; 2020. p. 325-56.
[25] Winter DA, Karolewska-Bochenek K, Lazowska-Przeorek I, Lionetti P, Mearin ML, Chong SK, et al. Pediatric IBD-unclassified Is Less Common than Previously Reported; Results of an 8-Year Audit of the EUROKIDS Registry. Inflamm Bowel Dis. 2015;21(9):2145-53.
[26] Thurgate LE, Lemberg DA, Day AS, Leach ST. An Overview of Inflammatory Bowel Disease Unclassified in Children. Inflammatory Intestinal Diseases. 2019;4(3):97-103.
[27] Rinawi F, Assa A, Eliakim R, Mozer-Glassberg Y, Nachmias Friedler V, Niv Y, et al. The natural history of pediatric-onset IBD-unclassified and prediction of Crohn’s disease reclassification: a 27-year study. Scandinavian Journal of Gastroenterology. 2017;52(5):558-63.
[28] Satsangi J, Silverberg MS, Vermeire S, Colombel JF. The Montreal classification of inflammatory bowel disease: controversies, consensus, and implications. Gut. 2006;55(6):749-53.
[29] Levine A, Griffiths A, Markowitz J, Wilson DC, Turner D, Russell RK, et al. Pediatric modification of the Montreal classification for inflammatory bowel disease: the Paris classification. Inflamm Bowel Dis. 2011;17(6):1314-21.
[30] Turner D, Otley AR, Mack D, Hyams J, de Bruijne J, Uusoue K, et al. Development, validation, and evaluation of a pediatric ulcerative colitis activity index: a prospective multicenter study. Gastroenterology. 2007;133(2):423-32.
[31] Hyams JS, Ferry GD, Mandel FS, Gryboski JD, Kibort PM, Kirschner BS, et al. Development and validation of a pediatric Crohn's disease activity index. J Pediatr Gastroenterol Nutr. 1991;12(4):439-47.
[32] Turner D, Griffiths AM, Walters TD, Seah T, Markowitz J, Pfefferkorn M, et al. Appraisal of the pediatric Crohn's disease activity index on four prospectively collected datasets: recommended cutoff values and clinimetric properties. The American journal of gastroenterology. 2010;105(9):2085-92.
[33] Sandhu BK, Fell JME, Beattie RM, Mitton SG, Wilson DC, Jenkins H, et al. Guidelines for the Management of Inflammatory Bowel Disease in Children in the United Kingdom. Journal of Pediatric Gastroenterology and Nutrition. 2010;50.
[34] Hanauer SB, Feagan BG, Lichtenstein GR, Mayer LF, Schreiber S, Colombel JF, et al. Maintenance infliximab for Crohn's disease: the ACCENT I randomised trial. Lancet (London, England). 2002;359(9317):1541-9.
[35] Schreiber S, Khaliq-Kareemi M, Lawrance IC, Thomsen OØ, Hanauer SB, McColm J, et al. Maintenance Therapy with Certolizumab Pegol for Crohn's Disease. New England Journal of Medicine. 2007;357(3):239-50.
[36] Hyams J, Crandall W, Kugathasan S, Griffiths A, Olson A, Johanns J, et al. Induction and maintenance infliximab therapy for the treatment of moderate-to-severe Crohn's disease in children. Gastroenterology. 2007;132(3):863-73; quiz 1165-6.
[37] Danese S, Fiorino G, Peyrin-Biroulet L. Early intervention in Crohn’s disease: towards disease modification trials. Gut. 2017;66(12):2179.

[38] Ruemmele FM, Veres G, Kolho KL, Griffiths A, Levine A, Escher JC, et al. Consensus guidelines of ECCO/ESPGHAN on the medical management of pediatric Crohn's disease. Journal of Crohn's and Colitis. 2014;8(10):1179-207.
[39] Aloi M, Nuti F, Stronati L, Cucchiara S. Advances in the medical management of paediatric IBD. Nature Reviews Gastroenterology & Hepatology. 2014;11(2):99-108.
[40] Ashton JJ, Ennis S, Beattie RM. Early-onset paediatric inflammatory bowel disease. The Lancet Child & adolescent health. 2017;1(2):147-58.
[41] Lucaciu LA, Seicean R, Seicean A. Small molecule drugs in the treatment of inflammatory bowel diseases: which one, when and why? - a systematic review. European journal of gastroenterology & hepatology. 2020;32(6):669-77.
[42] Nasiri S, Kuenzig ME, Benchimol EI. Long-term outcomes of pediatric inflammatory bowel disease. Seminars in Pediatric Surgery. 2017;26(6):398-404.
[43] Burisch J, Munkholm P. The epidemiology of inflammatory bowel disease. Scandinavian Journal of Gastroenterology. 2015;50(8):942-51.
[44] Frolkis AD, Dykeman J, Negrón ME, Debruyn J, Jette N, Fiest KM, et al. Risk of surgery for inflammatory bowel diseases has decreased over time: a systematic review and meta-analysis of population-based studies. Gastroenterology. 2013;145(5):996-1006.
[45] Fradet C, Kern J, Atanasov P, Wirth D, Borsi A. Impact of surgery and its complications in ulcerative colitis patients in clinical practice: A systematic literature review of real-world evidence in Europe. International Journal of Surgery Open. 2020;22:22-32.
[46] Axelrad JE, Olén O, Sachs MC, Erichsen R, Pedersen L, Halfvarson J, et al. Inflammatory bowel disease and risk of small bowel cancer: a binational population-based cohort study from Denmark and Sweden. Gut. 2020:gutjnl-2020-320945.
[47] Ott C, Schölmerich J. Extraintestinal manifestations and complications in IBD. Nature Reviews Gastroenterology & Hepatology. 2013;10(10):585-95.
[48] Nameirakpam J, Rikhi R, Rawat SS, Sharma J, Suri D. Genetics on early onset inflammatory bowel disease: An update. Genes & Diseases. 2020;7(1):93-106.
[49] Kelsen JR, Sullivan KE, Rabizadeh S, Singh N, Snapper S, Elkadri A, et al. NASPGHAN Position Paper on The Evaluation and Management for Patients with Very Early-Onset Inflammatory Bowel Disease (VEO-IBD). J Pediatr Gastroenterol Nutr. 2019.
[50] Kammermeier J, Dziubak R, Pescarin M, Drury S, Godwin H, Reeve K, et al. Phenotypic and Genotypic Characterisation of Inflammatory Bowel Disease Presenting Before the Age of 2 years. Journal of Crohn's and Colitis. 2017;11(1):60-9.
[51] Kotlarz D, Beier R, Murugan D, Diestelhorst J, Jensen O, Boztug K, et al. Loss of interleukin-10 signaling and infantile inflammatory bowel disease: implications for diagnosis and therapy. Gastroenterology. 2012;143(2):347-55.
[52] Gambineri E, Ciullini Mannurita S, Hagin D, Vignoli M, Anover-Sombke S, DeBoer S, et al. Clinical, Immunological, and Molecular Heterogeneity of 173 Patients With the Phenotype of Immune Dysregulation, Polyendocrinopathy, Enteropathy, X-Linked (IPEX) Syndrome. Frontiers in Immunology. 2018;9(2411).
[53] Guariso G, Gasparetto M, Visonà Dalla Pozza L, D'Incà R, Zancan L, Sturniolo G, et al. Inflammatory Bowel Disease Developing in Paediatric and Adult Age. Journal of Pediatric Gastroenterology and Nutrition. 2010;51(6).
[54] Ruemmele FM, Turner D. Differences in the management of pediatric and adult onset ulcerative colitis — lessons from the joint ECCO and ESPGHAN consensus guidelines for the management of pediatric ulcerative colitis. Journal of Crohn's and Colitis. 2014;8(1):1-4.
[55] Okou DT, Kugathasan S. Role of genetics in pediatric inflammatory bowel disease. Inflammatory bowel diseases. 2014;20(10):1878-84.
[56] Ostrowski J, Paziewska A, Lazowska I, Ambrozkiewicz F, Goryca K, Kulecka M, et al. Genetic architecture differences between pediatric and adult-onset inflammatory bowel diseases in the Polish population. Scientific Reports. 2016;6(1):39831.

[57] Crowley E, Warner N, Pan J, Khalouei S, Elkadri A, Fiedler K, et al. Prevalence and Clinical Features of Inflammatory Bowel Diseases Associated With Monogenic Variants, Identified by Whole-Exome Sequencing in 1000 Children at a Single Center. Gastroenterology. 2020;158(8):2208-20.
[58] Kelsen JR, Dawany N, Martinez A, Grochowski CM, Maurer K, Rappaport E, et al. A de novo whole gene deletion of XIAP detected by exome sequencing analysis in very early onset inflammatory bowel disease: A case report. BMC Gastroenterology. 2015;15 (1) (no pagination)(160).
[59] Tysk C, Lindberg E, Järnerot G, Flodérus-Myrhed B. Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. Gut. 1988;29(7):990-6.
[60] Brant SR. Update on the heritability of inflammatory bowel disease: The importance of twin studies. Inflammatory Bowel Diseases. 2011;17(1):1-5.
[61] Orholm M, Munkholm P, Langholz E, Nielsen OH, Sørensen TIA, Binder V. Familial Occurrence of Inflammatory Bowel Disease. New England Journal of Medicine. 1991;324(2):84-8.
[62] Xavier RJ, Rioux JD. Genome-wide association studies: a new window into immune-mediated diseases. Nature reviews Immunology. 2008;8(8):631-43.
[63] Jostins L, Ripke S, Weersma RK, Duerr RH, McGovern DP, Hui KY, et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 2012;491(7422):119-24.
[64] Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. Nature genetics. 2015;47(9):979-86.
[65] Huang H, Fang M, Jostins L, Umićević Mirkov M, Boucher G, Anderson CA, et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. Nature. 2017;547(7662):173-8.
[66] Franke A, McGovern DPB, Barrett JC, Wang K, Radford-Smith GL, Ahmad T, et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nature genetics. 2010;42(12):1118-25.
[67] Khor B, Gardet A, Xavier RJ. Genetics and pathogenesis of inflammatory bowel disease. Nature. 2011;474(7351):307-17.
[68] Gordon H, Trier Moller F, Andersen V, Harbord M. Heritability in inflammatory bowel disease: from the first twin study to genome-wide association studies. Inflamm Bowel Dis. 2015;21(6):1428-34.
[69] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57-74.
[70] Freeman JL, Perry GH, Feuk L, Redon R, McCarroll SA, Altshuler DM, et al. Copy number variation: new insights in genome diversity. Genome research. 2006;16(8):949-61.
[71] Hagemann IS. Chapter 1 - Overview of Technical Aspects and Chemistries of Next-Generation Sequencing. In: Kulkarni S, Pfeifer J, editors. Clinical Genomics. Boston: Academic Press; 2015. p. 3-19.
[72] Wetterstrand K. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) [Available from: www.genome.gov/sequencingcostsdata.
[73] Tsai EA, Shakbatyan R, Evans J, Rossetti P, Graham C, Sharma H, et al. Bioinformatics Workflow for Clinical Whole Genome Sequencing at Partners HealthCare Personalized Medicine. Journal of personalized medicine. 2016;6(1).
[74] Lam HYK, Clark MJ, Chen R, Chen R, Natsoulis G, O'Huallachain M, et al. Performance comparison of whole-genome sequencing platforms. Nature Biotechnology. 2012;30(1):78-82.
[75] Elliott DJ. Illuminating the Transcriptome through the Genome. Genes (Basel). 2014;5(1):235-53.
[76] Rabbani B, Tekin M, Mahdieh N. The promise of whole-exome sequencing in medical genetics. Journal of Human Genetics. 2014;59(1):5-15.
[77] Belkadi A, Bolze A, Itan Y, Cobat A, Vincent QB, Antipenko A, et al. Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. Proceedings of the National Academy of Sciences. 2015;112(17):5473.

[78] Poptsova MS, Il'icheva IA, Nechipurenko DY, Panchenko LA, Khodikov MV, Oparina NY, et al. Non-random DNA fragmentation in next-generation sequencing. Scientific Reports. 2014;4(1):4532.

[79] Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. Exome sequencing as a tool for Mendelian disease gene discovery. Nat Rev Genet. 2011;12(11):745-55.

[80] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754-60.

[81] Consortium GR.

[82] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.

[83] Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526(7571):68-74.

[84] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature methods. 2010;7(4):248-9.

[85] Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research. 2003;31(13):3812-4.

[86] Liu X, Wu C, Li C, Boerwinkle E. dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Human Mutation. 2016;37(3):235-41.

[87] Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLOS Computational Biology. 2010;6(12):e1001025.

[88] Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M. CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic acids research. 2018;47(D1):D886-D94.

[89] Quang D, Chen Y, Xie X. DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2014;31(5):761-3.

[90] Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research. 2014;42(Database issue):D980-D5.

[91] Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, et al. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat. 2003;21(6):577-81.

[92] Martin AR, Williams E, Foulger RE, Leigh S, Daugherty LC, Niblock O, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. Nature genetics. 2019;51(11):1560-5.

[93] Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine. 2015;17(5):405-23.

[94] Mueller WF, Larsen LSZ, Garibaldi A, Hatfield GW, Hertel KJ. The Silent Sway of Splicing by Synonymous Substitutions. J Biol Chem. 2015;290(46):27700-11.

[95] McCarthy DJ, Humburg P, Kanapin A, Rivas MA, Gaulton K, Cazier J-B, et al. Choice of transcripts and software has a large effect on variant annotation. Genome Medicine. 2014;6(3):26.

[96] Graham DB, Xavier RJ. Pathway paradigms revealed from the genetics of inflammatory bowel disease. Nature. 2020;578(7796):527-39.

[97] Geremia A, Biancheri P, Allan P, Corazza GR, Di Sabatino A. Innate and adaptive immunity in inflammatory bowel disease. Autoimmunity Reviews. 2014;13(1):3-10.

[98] Liu T, Zhang L, Joo D, Sun S-C. NF-κB signaling in inflammation. Signal Transduct Target Ther. 2017;2:17023.

[99] Kumaran Satyanarayanan S, El Kebir D, Soboh S, Butenko S, Sekheri M, Saadi J, et al. IFN-β is a macrophage-derived effector cytokine facilitating the resolution of bacterial inflammation. Nature Communications. 2019;10(1):3471.
[100] Wehkamp J, Harder J, Weichenthal M, Schwab M, Schäffeler E, Schlee M, et al. NOD2 (CARD15) mutations in Crohn's disease are associated with diminished mucosal alpha-defensin expression. Gut. 2004;53(11):1658-64.
[101] Watanabe T, Kitani A, Murray PJ, Strober W. NOD2 is a negative regulator of Toll-like receptor 2-mediated T helper type 1 responses. Nature immunology. 2004;5(8):800-8.
[102] Mukherjee T, Hovingh ES, Foerster EG, Abdel-Nour M, Philpott DJ, Girardin SE. NOD1 and NOD2 in inflammation, immunity and disease. Archives of Biochemistry and Biophysics. 2019;670:69-81.
[103] Nguyen GT, Green ER, Mecsas J. Neutrophils to the ROScue: Mechanisms of NADPH Oxidase Activation and Bacterial Resistance. Frontiers in Cellular and Infection Microbiology. 2017;7(373).
[104] Heyworth PG, Cross AR, Curnutte JT. Chronic granulomatous disease. Current Opinion in Immunology. 2003;15(5):578-84.
[105] Panday A, Sahoo MK, Osorio D, Batra S. NADPH oxidases: an overview from structure to innate immunity-associated pathologies. Cellular & molecular immunology. 2015;12(1):5-23.
[106] Khoshnevisan R, Anderson M, Babcock S, Anderson S, Illig D, Marquardt B, et al. NOX1 Regulates Collective and Planktonic Cell Migration: Insights From Patients With Pediatric-Onset IBD and NOX1 Deficiency. Inflammatory Bowel Diseases. 2020;26(8):1166-76.
[107] Schwerd T, Bryant RV, Pandey S, Capitani M, Meran L, Cazier JB, et al. NOX1 loss-of-function genetic variants in patients with inflammatory bowel disease. Mucosal immunology. 2018;11(2):562-74.
[108] Hayes P, Dhillon S, O'Neill K, Thoeni C, Hui KY, Elkadri A, et al. Defects in NADPH Oxidase Genes NOX1 and DUOX2 in Very Early Onset Inflammatory Bowel Disease. Cell Mol Gastroenterol Hepatol. 2015;1(5):489-502.
[109] Ashton JJ, Andreoletti G, Coelho T, Haggarty R, Batra A, Afzal NA, et al. Identification of Variants in Genes Associated with Single-gene Inflammatory Bowel Disease by Whole-exome Sequencing. Inflammatory Bowel Diseases. 2016;22(10):2317-27.
[110] Denson LA, Jurickova I, Karns R, Shaw KA, Cutler DJ, Okou DT, et al. Clinical and Genomic Correlates of Neutrophil Reactive Oxygen Species Production in Pediatric Patients With Crohn's Disease. Gastroenterology. 2018;154(8):2097-110.
[111] Eglinton TW, Roberts R, Pearson J, Barclay M, Merriman TR, Frizelle FA, et al. Clinical and genetic risk factors for perianal Crohn's disease in a population-based cohort. The American journal of gastroenterology. 2012;107(4):589-96.
[112] Ahluwalia B, Moraes L, Magnusson MK, Öhman L. Immunopathogenesis of inflammatory bowel disease and mechanisms of biological therapies. Scandinavian Journal of Gastroenterology. 2018;53(4):379-89.
[113] Uhlig HH, Charbit-Henrion F, Kotlarz D, Shouval DS, Schwerd T, Strisciuglio C, et al. Clinical Genomics for the Diagnosis of Monogenic Forms of Inflammatory Bowel Disease: A Position Paper From the Paediatric IBD Porto Group of European Society of Paediatric Gastroenterology, Hepatology and Nutrition. J Pediatr Gastroenterol Nutr. 2021;72(3):456-73.
[114] Sahin S, Adrovic A, Barut K, Ugurlu S, Turanli ET, Ozdogan H, et al. Clinical, imaging and genotypical features of three deceased and five surviving cases with ADA2 deficiency. Rheumatology international. 2018;38(1):129-36.
[115] Afzali B, Grönholm J, Vandrovcova J, O'Brien C, Sun HW, Vanderleyden I, et al. BACH2 immunodeficiency illustrates an association between super-enhancers and haploinsufficiency. Nature immunology. 2017;18(7):813-23.
[116] Mao L, Kitani A, Similuk M, Oler AJ, Albenberg L, Kelsen J, et al. Loss-of-function CARD8 mutation causes NLRP3 inflammasome activation and Crohn's disease. The Journal of clinical investigation. 2018;128(5):1793-806.
[117] Magg T, Shcherbina A, Arslan D, Desai MM, Wall S, Mitsialis V, et al. CARMIL2 Deficiency Presenting as Very Early Onset Inflammatory Bowel Disease. Inflamm Bowel Dis. 2019;25(11):1788-95.

[118] Lehle AS, Farin HF, Marquardt B, Michels BE, Magg T, Li Y, et al. Intestinal Inflammation and Dysregulated Immunity in Patients With Inherited Caspase-8 Deficiency. Gastroenterology. 2019;156(1):275-8.

[119] Ozen A. CHAPLE syndrome uncovers the primary role of complement in a familial form of Waldmann's disease. Immunological reviews. 2019;287(1):20-32.

[120] Zeissig S, Petersen BS, Tomczak M, Melum E, Huc-Claustre E, Dougan SK, et al. Early-onset Crohn's disease and autoimmunity associated with a variant in CTLA-4. Gut. 2015;64(12):1889-97.

[121] Arnadottir GA, Norddahl GL, Gudmundsdottir S, Agustsdottir AB, Sigurdsson S, Jensson BO, et al. A homozygous loss-of-function mutation leading to CYBC1 deficiency causes chronic granulomatous disease. Nat Commun. 2018;9(1):4447.

[122] Dobbs K, Domínguez Conde C, Zhang SY, Parolini S, Audry M, Chou J, et al. Inherited DOCK2 Deficiency in Patients with Early-Onset Invasive Infections. The New England journal of medicine. 2015;372(25):2409-22.

[123] Keller MD, Pandey R, Li D, Glessner J, Tian L, Henrickson SE, et al. Mutation in IRF2BP2 is responsible for a familial form of common variable immunodeficiency disorder. The Journal of allergy and clinical immunology. 2016;138(2):544-50.e4.

[124] Patel N, El Mouzan MI, Al-Mayouf SM, Adly N, Mohamed JY, Al Mofarreh MA, et al. Study of Mendelian forms of Crohn&#039;s disease in Saudi Arabia reveals novel risk loci and alleles. Gut. 2014;63(11):1831.

[125] Punwani D, Wang H, Chan AY, Cowan MJ, Mallott J, Sunderam U, et al. Combined immunodeficiency due to MALT1 mutations, treated by hematopoietic cell transplantation. Journal of clinical immunology. 2015;35(2):135-46.

[126] Boland BS, Widjaja CE, Banno A, Zhang B, Kim SH, Stoven S, et al. Immunodeficiency and autoimmune enterocolopathy linked to NFAT5 haploinsufficiency. Journal of immunology (Baltimore, Md : 1950). 2015;194(6):2551-60.

[127] Romberg N, Al Moussawi K, Nelson-Williams C, Stiegler AL, Loring E, Choi M, et al. Mutation of NLRC4 causes a syndrome of enterocolitis and autoinflammation. Nature genetics. 2014;46(10):1135-9.

[128] Schwerd T, Pandey S, Yang HT, Bagola K, Jameson E, Jung J, et al. Impaired antibacterial autophagy links granulomatous intestinal inflammation in Niemann-Pick disease type C1 and XIAP deficiency with NOD2 variants in Crohn's disease. Gut. 2017;66(6):1060-73.

[129] Brooke MA, Longhurst HJ, Plagnol V, Kirkby NS, Mitchell JA, Rüschendorf F, et al. Cryptogenic multifocal ulcerating stenosing enteritis associated with homozygous deletion mutations in cytosolic phospholipase A2-α. Gut. 2014;63(1):96-104.

[130] Starokadomskyy P, Gemelli T, Rios JJ, Xing C, Wang RC, Li H, et al. DNA polymerase-α regulates the activation of type I interferons through cytosolic RNA:DNA synthesis. Nature immunology. 2016;17(5):495-504.

[131] Brauer PM, Pessach IM, Clarke E, Rowe JH, Ott de Bruin L, Lee YN, et al. Modeling altered T-cell development with induced pluripotent stem cells from patients with RAG1-dependent immune deficiencies. Blood. 2016;128(6):783-93.

[132] Uhlig HH, Schwerd T, Koletzko S, Shah N, Kammermeier J, Elkadri A, et al. The Diagnostic Approach to Monogenic Very Early Onset Inflammatory Bowel Disease. Gastroenterology. 2014;147(5):990-1007.e3.

[133] Zhang N, Heruth DP, Wu W, Zhang LQ, Nsumu MN, Shortt K, et al. Functional characterization of SLC26A3 c.392C>G (p.P131R) mutation in intestinal barrier function using CRISPR/CAS9-created cell models. Cell & Bioscience. 2019;9(1):40.

[134] Flanagan SE, Haapaniemi E, Russell MA, Caswell R, Allen HL, De Franco E, et al. Activating germline mutations in STAT3 cause early-onset multi-organ autoimmune disease. Nature genetics. 2014;46(8):812-4.

[135] Kelsen JR, Ouahed J, Spessott WA, Kooshesh K, Sanmillan ML, Dawany N, et al. 25 MUTATIONS IN STXBP3 CONTRIBUTE TO VERY EARLY ONSET OF IBD, IMMUNODEFICIENCY AND HEARING LOSS. Gastroenterology. 2018;154(1, Supplement):S40-S1.
[136] Kotlarz D, Marquardt B, Baroy T, Lee WS, Konnikova L, Hollizeck S, et al. Human TGF-beta1 deficiency causes severe inflammatory bowel disease and encephalopathy. Nature genetics. 2018;50(3):344-8.
[137] Patel R, Coulter LL, Rimmer J, Parkes M, Chinnery PF, Swift O. Mitochondrial neurogastrointestinal encephalopathy: a clinicopathological mimic of Crohn's disease. BMC Gastroenterol. 2019;19(1):11.
[138] Uchiyama Y, Kim CA, Pastorino AC, Ceroni J, Lima PP, de Barros Dorna M, et al. Primary immunodeficiency with chronic enteropathy and developmental delay in a boy arising from a novel homozygous RIPK1 variant. Journal of Human Genetics. 2019;64(9):955-60.
[139] Salter CG, Cai Y, Lo B, Helman G, Taylor H, McCartney A, et al. Biallelic PI4KA variants cause neurological, intestinal and immunological disease. Brain. 2021;144(12):3597-610.
[140] Formankova R, Kanderova V, Rackova M, Svaton M, Brdicka T, Riha P, et al. Novel SAMD9 Mutation in a Patient With Immunodeficiency, Neutropenia, Impaired Anti-CMV Response, and Severe Gastrointestinal Involvement. Front Immunol. 2019;10:2194.
[141] Derakhshan D, Taherifard E, Taherifard E, Sajedianfard S, Derakhshan A. A novel frame shift mutation in STIM1 gene causing primary immunodeficiency. Intractable Rare Dis Res. 2020;9(2):109-12.
[142] Wang L, Aschenbrenner D, Zeng Z, Cao X, Mayr D, Mehta M, et al. Gain-of-function variants in SYK cause immune dysregulation and systemic inflammation in humans and mice. Nature genetics. 2021;53(4):500-10.
[143] Torkamani A, Wineinger NE, Topol EJ. The personal and clinical utility of polygenic risk scores. Nature Reviews Genetics. 2018;19(9):581-90.
[144] Chen G-B, Lee SH, Montgomery GW, Wray NR, Visscher PM, Gearry RB, et al. Performance of risk prediction for inflammatory bowel disease based on genotyping platform and genomic risk score method. BMC Medical Genetics. 2017;18(1):94.
[145] Vancamelbeke M, Vanuytsel T, Farré R, Verstockt S, Ferrante M, Van Assche G, et al. Genetic and Transcriptomic Bases of Intestinal Epithelial Barrier Dysfunction in Inflammatory Bowel Disease. Inflammatory bowel diseases. 2017;23(10):1718-29.
[146] Serra EG, Schwerd T, Moutsianas L, Cavounidis A, Fachal L, Pandey S, et al. Somatic mosaicism and common genetic variation contribute to the risk of very-early-onset inflammatory bowel disease. Nature Communications. 2020;11(1):995.
[147] Fatima M, Pasha M. Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications. 2017;09(01):1-16.
[148] James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning with Applications in R. 1 ed: Springer-Verlag New York; 2013. XIV, 426 p.
[149] Emura T, Matsui S, Chen H-Y. compound.Cox: Univariate feature selection and compound covariate for predicting survival. Computer Methods and Programs in Biomedicine. 2019;168:21-37.
[150] Lai C, Reinders MJT, Wessels L. Random subspace method for multivariate feature selection. Pattern Recognition Letters. 2006;27(10):1067-76.
[151] Guyon I, Weston J, Barnhill S, Vapnik V. Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning. 2002;46(1):389-422.
[152] Deng H, Runger G. Gene selection with guided regularized random forest. Pattern Recognition. 2013;46(12):3483-9.
[153] Jeni LA, Cohn JF, Torre FDL, editors. Facing Imbalanced Data--Recommendations for the Use of Performance Metrics. 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction; 2013 2-5 Sept. 2013.
[154] Cooper GS, Bynum ML, Somers EC. Recent insights in the epidemiology of autoimmune diseases: improved prevalence estimates and understanding of clustering of diseases. J Autoimmun. 2009;33(3-4):197-207.

[155] Hayter SM, Cook MC. Updated assessment of the prevalence, spectrum and case definition of autoimmune disease. Autoimmun Rev. 2012;11(10):754-65.
[156] Eaton WW, Rose NR, Kalaydjian A, Pedersen MG, Mortensen PB. Epidemiology of autoimmune diseases in Denmark. J Autoimmun. 2007;29(1):1-9.
[157] Male DK, Roitt IM, Roth DB, Roitt IM. Immunology. Eighth ed: Saunders; 2013.
[158] Tobón GJ, Pers J-O, Cañas CA, Rojas-Villarraga A, Youinou P, Anaya J-M. Are autoimmune diseases predictable? Autoimmunity Reviews. 2012;11(4):259-66.
[159] Larizza D, Calcaterra V, Klersy C, Badulli C, Caramagna C, Ricci A, et al. Common immunogenetic profile in children with multiple autoimmune diseases: the signature of HLA-DQ pleiotropic genes. Autoimmunity. 2012;45(6):470-5.
[160] Goodnow CC, Sprent J, de St Groth BF, Vinuesa CG. Cellular and genetic mechanisms of self tolerance and autoimmunity. Nature. 2005;435(7042):590-7.
[161] Bizzaro N. Autoantibodies as predictors of disease: The clinical and experimental evidence. Autoimmunity Reviews. 2007;6(6):325-33.
[162] Kuchroo VK, Ohashi PS, Sartor RB, Vinuesa CG. Dysregulation of immune homeostasis in autoimmune diseases. Nat Med. 2012;18(1):42-7.
[163] Costenbader KH, Gay S, Alarcón-Riquelme ME, Iaccarino L, Doria A. Genes, epigenetic regulation and environmental factors: Which is the most relevant in developing autoimmune diseases? Autoimmunity Reviews. 2012;11(8):604-9.
[164] Hewagama A, Richardson B. The genetics and epigenetics of autoimmune diseases. Journal of Autoimmunity. 2009;33(1):3-11.
[165] Aslani S, Mahmoudi M, Karami J, Jamshidi AR, Malekshahi Z, Nicknam MH. Epigenetic alterations underlying autoimmune diseases. Autoimmunity. 2016;49(2):69-83.
[166] Gerber DE. Targeted therapies: a new generation of cancer treatments. Am Fam Physician. 2008;77(3):311-9.
[167] Cho JH, Feldman M. Heterogeneity of autoimmune diseases: pathophysiologic insights from genetics and implications for new therapies. Nature Medicine. 2015;21:730.
[168] Simon TA, Kawabata H, Ray N, Baheti A, Suissa S, Esdaile JM. Prevalence of Co-existing Autoimmune Disease in Rheumatoid Arthritis: A Cross-Sectional Study. Adv Ther. 2017;34(11):2481-90.
[169] Gilhus NE, Nacu A, Andersen JB, Owe JF. Myasthenia gravis and risks for comorbidity. European Journal of Neurology. 2015;22(1):17-23.
[170] Ruggeri RM, Trimarchi F, Giuffrida G, Certo R, Cama E, Campennì A, et al. Autoimmune comorbidities in Hashimoto’s thyroiditis: different patterns of association in adulthood and childhood/adolescence. 2017;176(2):133.
[171] Gill L, Zarbo A, Isedeh P, Jacobsen G, Lim HW, Hamzavi I. Comorbid autoimmune diseases in patients with vitiligo: A cross-sectional study. Journal of the American Academy of Dermatology. 2016;74(2):295-302.
[172] Jones M, Koziel C, Larsen D, Berry P, Kubatka-Willms E. Progress in the Enhanced Use of Electronic Medical Records: Data From the Ontario Experience. JMIR Med Inform. 2017;5(1):e5-e.
[173] Teschendorff AE. Avoiding common pitfalls in machine learning omic data science. Nature Materials. 2019;18(5):422-7.
[174] Manzoni C, Kia DA, Vandrovcova J, Hardy J, Wood NW, Lewis PA, et al. Genome, transcriptome and proteome: the rise of omics data and their integration in biomedical sciences. Briefings in Bioinformatics. 2016;19(2):286-302.
[175] Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Medical Genomics. 2015;8(1):33.
[176] Jiang F, Jiang Y, Zhi H, Dong Y, Li H, Ma S, et al. Artificial intelligence in healthcare: past, present and future. Stroke Vasc Neurol. 2017;2(4):230-43.
[177] Moher D, Liberati A, Tetzlaff J, Altman DG, The PG. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLOS Medicine. 2009;6(7):e1000097.

[178] Team RC. R: A language and environment for statistical computing. 2019.
[179] Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer-Verlag New York; 2016.
[180] Briggs FBS, Yu JC, Davis MF, Jiangyang J, Fu S, Parrotta E, et al. Multiple sclerosis risk factors contribute to onset heterogeneity. Multiple Sclerosis and Related Disorders. 2019;28:11-6.
[181] Ahmadi A, Davoudi S, Daliri MR. Computer Aided Diagnosis System for multiple sclerosis disease based on phase to amplitude coupling in covert visual attention. Comput Methods Programs Biomed. 2019;169:9-18.
[182] Zhang H, Alberts E, Pongratz V, Mühlau M, Zimmer C, Wiestler B, et al. Predicting conversion from clinically isolated syndrome to multiple sclerosis–An imaging-based machine learning approach. NeuroImage: Clinical. 2019;21:101593.
[183] Zurita M, Montalba C, Labbé T, Cruz JP, Dalboni da Rocha J, Tejos C, et al. Characterization of relapsing-remitting multiple sclerosis patients using support vector machine classifications of functional and diffusion MRI data. NeuroImage: Clinical. 2018;20:724-30.
[184] Wang S-H, Tang C, Sun J, Yang J, Huang C, Phillips P, et al. Multiple Sclerosis Identification by 14-Layer Convolutional Neural Network With Batch Normalization, Dropout, and Stochastic Pooling. Frontiers in Neuroscience. 2018;12(818).
[185] Neeb H, Schenk J. Multivariate prediction of multiple sclerosis using robust quantitative MR-based image metrics. Zeitschrift für Medizinische Physik. 2018.
[186] Lotsch J, Schiffmann S, Schmitz K, Brunkhorst R, Lerch F, Ferreiros N, et al. Machine-learning based lipid mediator serum concentration patterns allow identification of multiple sclerosis patients with high accuracy. Scientific Reports. 2018;8(1):14884.
[187] Tacchella A, Romano S, Ferraldeschi M, Salvetti M, Zaccaria A, Crisanti A, et al. Collaboration between a human group and artificial intelligence can improve prediction of multiple sclerosis course: a proof-of-principle study. F1000Research. 2017;6:2172.
[188] Lopez C, Tucker S, Salameh T, Tucker C. An unsupervised machine learning method for discovering patient clusters based on genetic signatures. Journal of Biomedical Informatics. 2018;85:30-9.
[189] Supratak A, Datta G, Gafson AR, Nicholas R, Guo Y, Matthews PM. Remote Monitoring in the Home Validates Clinical Gait Measures for Multiple Sclerosis. Frontiers in Neurology. 2018;9(561).
[190] Saccà V, Sarica A, Novellino F, Barone S, Tallarico T, Filippelli E, et al. Evaluation of machine learning algorithms performance for the prediction of early multiple sclerosis from resting-state FMRI connectivity data. Brain Imaging and Behavior. 2018.
[191] Mowry EM, Hedström AK, Gianfrancesco MA, Shao X, Schaefer CA, Shen L, et al. Incorporating machine learning approaches to assess putative environmental risk factors for multiple sclerosis. Multiple Sclerosis and Related Disorders. 2018;24:135-41.
[192] Yoo Y, Tang LYW, Brosch T, Li DKB, Kolind S, Vavasour I, et al. Deep learning of joint myelin and T1w MRI features in normal-appearing brain tissue to distinguish between multiple sclerosis patients and healthy controls. NeuroImage: Clinical. 2018;17:169-78.
[193] Kiiski H, Jollans L, Donnchadha SÓ, Nolan H, Lonergan R, Kelly S, et al. Machine Learning EEG to Predict Cognitive Functioning and Processing Speed Over a 2-Year Period in Multiple Sclerosis Patients and Controls. Brain Topography. 2018;31(3):346-63.
[194] Fiorini S, Verri A, Tacchino A, Ponzio M, Brichetto G, Barla A. A machine learning pipeline for multiple sclerosis course detection from clinical scales and patient reported outcomes. Conf Proc IEEE Eng Med Biol Soc. 2015;2015:4443-46.
[195] Zhong J, Chen DQ, Nantes JC, Holmes SA, Hodaie M, Koski L. Combined structural and functional patterns discriminating upper limb motor disability in multiple sclerosis using multivariate approaches. Brain Imaging Behav. 2017;11(3):754-68.
[196] Lötsch J, Thrun M, Lerch F, Brunkhorst R, Schiffmann S, Thomas D, et al. Machine-Learned Data Structures of Lipid Marker Serum Concentrations in Multiple Sclerosis Patients Differ from Those in Healthy Subjects. International journal of molecular sciences. 2017;18(6):1217.
[197] Karaca Y, Zhang YD, Cattani C, Ayan U. The Differential Diagnosis of Multiple Sclerosis Using Convex Combination of Infinite Kernels. CNS Neurol Disord Drug Targets. 2017;16(1):36-43.

[198] Ostmeyer J, Christley S, Rounds WH, Toby I, Greenberg BM, Monson NL, et al. Statistical classifiers for diagnosing disease from immune repertoires: a case study using multiple sclerosis. BMC Bioinformatics. 2017;18(1):401.

[199] McGinnis RS, Mahadevan N, Moon Y, Seagers K, Sheth N, Wright JA, Jr., et al. A machine learning approach for gait speed estimation using skin-mounted wearable sensors: From healthy controls to individuals with multiple sclerosis. PLoS One. 2017;12(6):e0178366.

[200] Zhao Y, Healy BC, Rotstein D, Guttmann CRG, Bakshi R, Weiner HL, et al. Exploration of machine learning techniques in predicting multiple sclerosis disease course. PLOS ONE. 2017;12(4):e0174866.

[201] Ion-Mărgineanu A, Kocevar G, Stamile C, Sima DM, Durand-Dubief F, Van Huffel S, et al. Machine Learning Approach for Classifying Multiple Sclerosis Courses by Combining Clinical Data with Lesion Loads and Magnetic Resonance Metabolic Features. Frontiers in Neuroscience. 2017;11(398).

[202] Kocevar G, Stamile C, Hannoun S, Cotton F, Vukusic S, Durand-Dubief F, et al. Graph Theory-Based Brain Connectivity for Automatic Classification of Multiple Sclerosis Clinical Courses. Frontiers in Neuroscience. 2016;10(478).

[203] Kosa P, Ghazali D, Tanigawa M, Barbour C, Cortese I, Kelley W, et al. Development of a Sensitive Outcome for Economical Drug Screening for Progressive Multiple Sclerosis Treatment. Frontiers in Neurology. 2016;7(131).

[204] Baranzini SE, Madireddy LR, Cromer A, D'Antonio M, Lehr L, Beelke M, et al. Prognostic biomarkers of IFNb therapy in multiple sclerosis patients. Multiple sclerosis (Houndmills, Basingstoke, England). 2015;21(7):894-904.

[205] Wottschel V, Alexander DC, Kwok PP, Chard DT, Stromillo ML, De Stefano N, et al. Predicting outcome in clinically isolated syndrome using machine learning. Neuroimage Clin. 2015;7:281-7.

[206] Crimi A, Commowick O, Maarouf A, Ferre JC, Bannier E, Tourbah A, et al. Predictive value of imaging markers at multiple sclerosis disease onset based on gadolinium- and USPIO-enhanced MRI and machine learning. PLoS One. 2014;9(4):e93024.

[207] Sweeney EM, Vogelstein JT, Cuzzocreo JL, Calabresi PA, Reich DS, Crainiceanu CM, et al. A Comparison of Supervised Machine Learning Algorithms and Feature Vectors for MS Lesion Segmentation Using Multimodal Structural MRI. PLOS ONE. 2014;9(4):e95753.

[208] Taschler B, Ge T, Bendfeldt K, Müller-Lenke N, Johnson TD, Nichols TE, editors. Spatial Modeling of Multiple Sclerosis for Disease Subtype Prediction2014; Cham: Springer International Publishing.

[209] Alaqtash M, Sarkodie-Gyan T, Yu H, Fuentes O, Brower R, Abdelgawad A. Automatic classification of pathological gait patterns using ground reaction forces and machine learning algorithms. Conf Proc IEEE Eng Med Biol Soc. 2011;2011:453-7.

[210] Goldstein BA, Hubbard AE, Cutler A, Barcellos LF. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. BMC genetics. 2010;11:49.

[211] Corvol JC, Pelletier D, Henry RG, Caillier SJ, Wang J, Pappas D, et al. Abrogation of T cell quiescence characterizes patients at high risk for multiple sclerosis after the initial neurological event. Proceedings of the National Academy of Sciences of the United States of America. 2008;105(33):11839-44.

[212] Briggs FB, Bartlett SE, Goldstein BA, Wang J, McCauley JL, Zuvich RL, et al. Evidence for CRHR1 in multiple sclerosis using supervised machine learning and meta-analysis in 12,566 individuals. Human molecular genetics. 2010;19(21):4286-95.

[213] Commowick O, Istace A, Kain M, Laurent B, Leray F, Simon M, et al. Objective Evaluation of Multiple Sclerosis Lesion Segmentation using a Data Management and Processing Infrastructure. Scientific Reports. 2018;8(1):13650.

[214] Ohanian D, Brown A, Sunnquist M, Furst J, Nicholson L, Klebek L, et al. Identifying Key Symptoms Differentiating Myalgic Encephalomyelitis and Chronic Fatigue Syndrome from Multiple Sclerosis. Neurology (E-Cronicon). 2016;4(2):41-5.
[215] Salem M, Cabezas M, Valverde S, Pareto D, Oliver A, Salvi J, et al. A supervised framework with intensity subtraction and deformation field features for the detection of new T2-w lesions in multiple sclerosis. NeuroImage: Clinical. 2018;17:607-15.
[216] Cabezas M, Oliver A, Valverde S, Beltran B, Freixenet J, Vilanova JC, et al. BOOST: A supervised approach for multiple sclerosis lesion segmentation. Journal of Neuroscience Methods. 2014;237:108-17.
[217] Zhang Y, Lu S, Zhou X, Yang M, Wu L, Liu B, et al. Comparison of machine learning methods for stationary wavelet entropy-based multiple sclerosis detection: decision tree, k-nearest neighbors, and support vector machine. Simulation. 2016;92(9):861-71.
[218] Birenbaum A, Greenspan H. Multi-view longitudinal CNN for multiple sclerosis lesion segmentation. Engineering Applications of Artificial Intelligence. 2017;65:111-8.
[219] Morrison C, Huckvale K, Corish B, Dorn J, Kontschieder P, O'Hara K, et al. Assessing Multiple Sclerosis With Kinect: Designing Computer Vision Systems for Real-World Use. Human-Computer Interaction. 2016;31(3/4):191-226.
[220] Liu J, Brodley CE, Healy BC, Chitnis T. Removing confounding factors via constraint-based clustering: An application to finding homogeneous groups of multiple sclerosis patients. Artificial Intelligence in Medicine. 2015;65(2):79-88.
[221] Chin CY, Hsieh SY, Tseng VS. EDram: Effective early disease risk assessment with matrix factorization on a large-scale medical database: A case study on rheumatoid arthritis. PLoS ONE. 2018;13 (11) (e0207579).
[222] Chocholova E, Bertok T, Jane E, Lorencova L, Holazova A, Belicka L, et al. Glycomics meets artificial intelligence - Potential of glycan analysis for identification of seropositive and seronegative rheumatoid arthritis patients revealed. Clinica Chimica Acta. 2018;481:49-55.
[223] Wu H, Cai L, Li D, Wang X, Zhao S, Zou F, et al. Metagenomics Biomarkers Selected for Prediction of Three Different Diseases in Chinese Population. BioMed Research International. 2018;2018 (2936257).
[224] Joo YB, Kim Y, Park Y, Kim K, Ryu JA, Lee S, et al. Biological function integrated prediction of severe radiographic progression in rheumatoid arthritis: A nested case control study. Arthritis and Rheumatology Conference: American College of Rheumatology/Association of Rheumatology Health Professionals Annual Scientific Meeting, ACR/ARHP. 2017;19(1):244.
[225] Andreu-Perez J, Garcia-Gancedo L, McKinnell J, Van der Drift A, Powell A, Hamy V, et al. Developing Fine-Grained Actigraphies for Rheumatoid Arthritis Patients from a Single Accelerometer Using Machine Learning. Sensors. 2017;17(9):2113.
[226] Orange DE, Agius P, DiCarlo EF, Robine N, Geiger H, Szymonifka J, et al. Identification of Three Rheumatoid Arthritis Disease Subtypes by Machine Learning Integration of Synovial Histologic Features and RNA Sequencing Data. Arthritis and Rheumatology. 2018;70(5):690-701.
[227] Ahmed U, Anwar A, Savage RS, Thornalley PJ, Rabbani N. Protein oxidation, nitration and glycation biomarkers for early-stage diagnosis of osteoarthritis of the knee and typing and progression of arthritic disease. Arthritis Research & Therapy. 2016;18(1):250.
[228] Miyoshi F, Honne K, Minota S, Okada M, Ogawa N, Mimura T. A novel method predicting clinical response using only background clinical data in RA patients before treatment with infliximab. Modern Rheumatology. 2016;26(6):813-6.
[229] Yeo L, Adlard N, Biehl M, Juarez M, Smallie T, Snow M, et al. Expression of chemokines CXCL4 and CXCL7 by synovial macrophages defines an early stage of rheumatoid arthritis. Annals of the Rheumatic Diseases. 2016;75(4):763-71.
[230] Zhou SM, Fernandez-Gutierrez F, Kennedy J, Cooksey R, Atkinson M, Denaxas S, et al. Defining Disease Phenotypes in Primary Care Electronic Health Records by a Machine Learning Approach: A Case Study in Identifying Rheumatoid Arthritis. PLoS ONE [Electronic Resource]. 2016;11(5):e0154515.

[231] Lin C, Karlson EW, Dligach D, Ramirez MP, Miller TA, Mo H, et al. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. J Am Med Inform Assoc. 2015;22(e1):e151-e61.
[232] Chen Y, Carroll RJ, Hinz ERM, Shah A, Eyler AE, Denny JC, et al. Applying active learning to high-throughput phenotyping algorithms for electronic health records data. J Am Med Inform Assoc. 2013;20:e253-e9.
[233] Lin C, Karlson EW, Canhao H, Miller TA, Dligach D, Chen PJ, et al. Automatic Prediction of Rheumatoid Arthritis Disease Activity from the Electronic Medical Records. PLoS ONE. 2013;8 (8) (e69932).
[234] Negi S, Juyal G, Senapati S, Prasad P, Gupta A, Singh S, et al. A genome-wide association study reveals ARL15, a novel non-HLA susceptibility gene for rheumatoid arthritis in North Indians. Arthritis and Rheumatism. 2013;65(12):3026-35.
[235] Pratt AG, Swan DC, Richardson S, Wilson G, Hilkens CMU, Young DA, et al. A CD4 T cell gene signature for early rheumatoid arthritis implicates interleukin 6-mediated STAT3 signalling, particularly in anti-citrullinated peptide antibody-negative disease. Annals of the Rheumatic Diseases. 2012;71(8):1374-81.
[236] Singh S, Kumar A, Panneerselvam K, Vennila JJ. Diagnosis of arthritis through fuzzy inference system. Journal of Medical Systems. 2012;36(3):1459-68.
[237] Kruppa J, Ziegler A, Konig IR. Risk estimation and risk prediction using machine-learning methods. Human Genetics. 2012;131(10):1639-54.
[238] Liu C, Ackerman HH, Carulli JP. A genome-wide screen of gene-gene interactions for rheumatoid arthritis susceptibility. Human Genetics. 2011;129(5):473-85.
[239] Nair SS, French RM, Laroche D, Thomas E. The Application of Machine Learning Algorithms to the Analysis of Electromyographic Patterns From Arthritic Patients. IEEE Trans Neural Syst Rehabil Eng. 2010;18(2):174-84.
[240] Briggs FBS, Ramsay PP, Madden E, Norris JM, Holers VM, Mikuls TR, et al. Supervised machine learning and logistic regression identifies novel epistatic risk factors with PTPN22 for rheumatoid arthritis. Genes and Immunity. 2010;11(3):199-208.
[241] Niu Q, Huang Z, Shi Y, Wang L, Pan X, Hu C. Specific serum protein biomarkers of rheumatoid arthritis detected by MALDI-TOF-MS combined with magnetic beads. International Immunology. 2010;22(7):611-8.
[242] Geurts P, Fillet M, de Seny D, Meuwis MA, Malaise M, Merville MP, et al. Proteomic mass spectra classification using decision tree based ensemble methods. Bioinformatics. 2005;21(14):3138-45.
[243] De Seny D, Fillet M, Meuwis MA, Geurts P, Lutteri L, Ribbens C, et al. Discovery of new rheumatoid arthritis biomarkers using the surface-enhanced laser desorption/ionization time-of-flight mass spectrometry proteinchip approach. Arthritis and Rheumatism. 2005;52(12):3801-12.
[244] Scheel AK, Netz UJ, Hermann KGA, Hielscher AH, Klose AD, Tresp V, et al. Laser Imaging Techniques for Follow-up Analysis of Joint Inflammation in Patients with Rheumatoid Arthritis. Medical Laser Application. 2003;18(3):198-205.
[245] Gronsbell J, Minnier J, Yu S, Liao K, Cai T. Automated Feature Selection of Predictors in Electronic Medical Records Data. Biometrics. 2018.
[246] Gossec L, Guyard F, Leroy D, Lafargue T, Seiler M, Jacquemin C, et al. Detection of flares by decrease in physical activity, collected using wearable activity trackers, in rheumatoid arthritis or axial spondyloarthritis: an application of Machine-Learning analyses in rheumatology. Arthritis Care Res (Hoboken). 2018;22:22.
[247] Lezcano-Valverde JM, Salazar F, Leon L, Toledano E, Jover JA, Fernandez-Gutierrez B, et al. Development and validation of a multivariate predictive model for rheumatoid arthritis mortality using a machine learning approach. Scientific Reports. 2017;7(1):10189.
[248] Gonzalez-Recio O, de Maturana EL, Vega AT, Engelman CD, Broman KW. Detecting single-nucleotide polymorphism by single-nucleotide polymorphism interactions in rheumatoid arthritis

using a two-step approach with machine learning and a Bayesian threshold least absolute shrinkage and selection operator (LASSO) model. BMC Proc. 2009;3 Suppl 7:S63.
[249] Heard BJ, Rosvold JM, Fritzler MJ, El-Gabalawy H, Wiley JP, Krawetz RJ. A computational method to differentiate normal individuals, osteoarthritis and rheumatoid arthritis patients using serum biomarkers. J R Soc Interface. 2014;11(97):20140428.
[250] Gronsbell JL, Cai T. Semi‐supervised approaches to efficient evaluation of model prediction performance. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2018;80(3):579-94.
[251] Van Looy S, Vander Cruyssen B, Meeus J, Wyns B, Westhovens R, Durez P, et al. Prediction of dose escalation for rheumatoid arthritis patients under infliximab treatment. Engineering Applications of Artificial Intelligence. 2006;19(7):819-28.
[252] Wyns B, Boullart L, Sette S, Baeten D, Hoffman I, De Keyser F. Prediction of arthritis using a modified Kohonen mapping and case based reasoning. Engineering Applications of Artificial Intelligence. 2004;17(2):205.
[253] Waljee AK, Lipson R, Wiitala WL, Zhang Y, Liu B, Zhu J, et al. Predicting Hospitalization and Outpatient Corticosteroid Use in Inflammatory Bowel Disease Patients Using Machine Learning. Inflammatory Bowel Diseases. 2018;24(1):45-53.
[254] Mossotto E, Ashton JJ, Coelho T, Beattie RM, MacArthur BD, Ennis S. Classification of Paediatric Inflammatory Bowel Disease using Machine Learning. Scientific reports. 2017;7(1):2427.
[255] Maeda Y, Kudo SE, Mori Y, Misawa M, Ogata N, Sasanuma S, et al. Fully automated diagnostic system with artificial intelligence using endocytoscopy to identify the presence of histologic inflammation associated with ulcerative colitis (with video). Gastrointestinal Endoscopy. 2018;89(2):408-15.
[256] Douglas GM, Hansen R, Jones CMA, Dunn KA, Comeau AM, Bielawski JP, et al. Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. Microbiome.6(1):13.
[257] Jain S, Kedia S, Sethi T, Bopanna S, Yadav DP, Goyal S, et al. Predictors of long-term outcomes in patients with acute severe colitis: A northern Indian cohort study. Journal of Gastroenterology and Hepatology (Australia). 2018;33(3):615-22.
[258] Waljee AK, Sauder K, Patel A, Segar S, Liu B, Zhang Y, et al. Machine learning algorithms for objective remission and clinical outcomes with thiopurines. Journal of Crohn's and Colitis. 2017;11(7):801-10.
[259] Isakov O, Dotan I, Ben-Shachar S. Machine Learning-Based Gene Prioritization Identifies Novel Candidate Risk Genes for Inflammatory Bowel Disease. Inflammatory Bowel Diseases. 2017;23(9):1516-23.
[260] Kang T, Ding W, Zhang L, Ziemek D, Zarringhalam K. A biological network-based regularized artificial neural network model for robust phenotype prediction from gene expression data. BMC Bioinformatics. 2017;18(1):565.
[261] Waljee AK, Liu B, Sauder K, Zhu J, Govani SM, Stidham RW, et al. Predicting corticosteroid-free endoscopic remission with vedolizumab in ulcerative colitis. Alimentary Pharmacology and Therapeutics. 2018;47(6):763-72.
[262] Eck A, Zintgraf LM, de Groot EFJ, de Meij TGJ, Cohen TS, Savelkoul PHM, et al. Interpretation of microbiota-based diagnostics by explaining individual classifier decisions. BMC Bioinformatics. 2017;18(1):441.
[263] Menti E, Lanera C, Lorenzoni G, Giachino DF, Marchi M, Gregori D, et al. Bayesian Machine Learning Techniques for revealing complex interactions among genetic and clinical factors in association with extra-intestinal Manifestations in IBD patients. Amia 2016;Annual Symposium proceedings. AMIA Symposium. 2016:884-93.
[264] Hubenthal M, Hemmrich-Stanisak G, Degenhardt F, Szymczak S, Du Z, Elsharawy A, et al. Sparse modeling reveals miRNA signatures for diagnostics of inflammatory bowel disease. PLoS ONE. 2015;10 (10) (e140155).
[265] Niehaus KE, Uhlig HH, Clifton DA. Phenotypic characterisation of Crohn's disease severity. Conference proceedings : . 2015;Annual International Conference of the IEEE Engineering in

Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference. 2015:7023-6.
[266] Wei Z, Wang W, Bradfield J, Li J, Cardinale C, Frackelton E, et al. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. American Journal of Human Genetics. 2013;92(6):1008-12.
[267] Cui H, Zhang X. Alignment-free supervised classification of metagenomes by recursive SVM. BMC Genomics. 2013;14 (1) (641).
[268] Waljee AK, Joyce JC, Wang S, Saxena A, Hart M, Zhu J, et al. Algorithms Outperform Metabolite Tests in Predicting Response of Patients With Inflammatory Bowel Disease to Thiopurines. Clinical Gastroenterology and Hepatology. 2010;8(2):143-50.
[269] Firouzi F, Rashidi M, Hashemi S, Kangavari M, Bahari A, Daryani NE, et al. A decision tree-based approach for determining low bone mineral density in inflammatory bowel disease using WEKA software. European Journal of Gastroenterology and Hepatology. 2007;19(12):1075-81.
[270] Ozawa T, Ishihara S, Fujishiro M, Saito H, Kumagai Y, Shichijo S, et al. Novel Computer-assisted Diagnosis System for Endoscopic Disease Activity in Patients with Ulcerative Colitis. Gastrointestinal Endoscopy. 2018.
[271] Reddy BK, Delen D, Agrawal RK. Predicting and explaining inflammation in Crohn's disease patients using predictive analytics methods and electronic medical record data. Health Inform J. 2018:1460458217751015.
[272] Forbes JD, Chen CY, Knox NC, Marrie RA, El-Gabalawy H, de Kievit T, et al. A comparative study of the gut microbiota in immune-mediated inflammatory diseases-does a common dysbiosis exist? Microbiome. 2018;6(1):221.
[273] Doherty MK, Ding T, Koumpouras C, Telesco SE, Monast C, Das A, et al. Fecal Microbiota Signatures Are Associated with Response to Ustekinumab Therapy among Crohn's Disease Patients. mBio. 2018;9(2):e02120-17.
[274] Han L, Maciejewski M, Gordon W, Afzelius L, Brockel C, Snapper SB, et al. A probabilistic pathway score (PROPS) for classification with applications to inflammatory bowel disease. Bioinformatics. 2018;34(6):985-93.
[275] Daneshjou R, Wang Y, Bromberg Y, Bovo S, Martelli PL, Babbi G, et al. Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. Human Mutation. 2017;38(9):1182-92.
[276] Giollo M, Jones DT, Carraro M, Leonardi E, Ferrari C, Tosatto SCE. Crohn disease risk prediction-Best practices and pitfalls with exome data. Human Mutation. 2017;38(9):1193-200.
[277] Yu S, Chakrabortty A, Liao KP, Cai T, Ananthakrishnan AN, Gainer VS, et al. Surrogate-assisted feature extraction for high-throughput phenotyping. J Am Med Inform Assoc. 2017;24(e1):e143-e9.
[278] Wisittipanit N, Rangwala H, Sikaroodi M, Keshavarzian A, Mutlu EA, Gillevet P. Classification methods for the analysis of LH-PCR data associated with inflammatory bowel disease patients. Int J Bioinform Res Appl. 2015;11(2):111-29.
[279] Ahmed S, Dey N, Ashour A, Sifaki-Pistolla D, Bălas-Timar D, Balas V, et al. Effect of fuzzy partitioning in Crohn's disease classification: a neuro-fuzzy-based approach. Med Biol Eng Comput. 2017;55(1):101-15.
[280] Mahapatra D, Vos FM, Buhmann JM. Active learning based segmentation of Crohns disease from abdominal MRI. Comput Methods Programs Biomed. 2016;128:75-85.
[281] Mahapatra D. Combining multiple expert annotations using semi-supervised learning and graph cuts for medical image segmentation. Computer Vision & Image Understanding. 2016;151:114-23.
[282] Pal LR, Kundu K, Yin Y, Moult J. CAGI4 Crohn's exome challenge: Marker SNP versus exome variant models for assigning risk of Crohn disease. Human Mutation. 2017;38(9):1225-34.
[283] Stawiski K, Pietrzak I, Mlynarski W, Fendler W, Szadkowska A. NIRCa: An artificial neural network-based insulin resistance calculator. Pediatric Diabetes. 2018;19(2):231-5.

[284] Ben Ali J, Hamdi T, Fnaiech N, Di Costanzo V, Fnaiech F, Ginoux JM. Continuous blood glucose level prediction of Type 1 Diabetes based on Artificial Neural Network. Biocybernetics and Biomedical Engineering. 2018;38(4):828-40.
[285] Maulucci G, Cordelli E, Rizzi A, De Leva F, Papi M, Ciasca G, et al. Phase separation of the plasma membrane in human red blood cells as a potential tool for diagnosis and progression monitoring of type 1 diabetes mellitus. PLoS ONE. 2017;12 (9) (e0184109).
[286] Siegel AP, Daneshkhah A, Hardin DS, Shrestha S, Varahramyan K, Agarwal M. Analyzing breath samples of hypoglycemic events in type 1 diabetes patients: Towards developing an alternative to diabetes alert dogs. Journal of Breath Research. 2017;11 (2) (026007).
[287] Zhao LP, Bolouri H, Zhao M, Geraghty DE, Lernmark A. An Object-Oriented Regression for Building Disease Predictive Models with Multiallelic HLA Genes. Genetic Epidemiology. 2016;40(4):315-32.
[288] Georga EI, Protopappas VC, Polyzos D, Fotiadis DI. Online prediction of glucose concentration in type 1 diabetes using extreme learning machines. Conference proceedings : . 2015;Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Conference. 2015:3262-5.
[289] Georga EI, Protopappas VC, Ardigo D, Polyzos D, Fotiadis DI. A Glucose Model Based on Support Vector Regression for the Prediction of Hypoglycemic Events Under Free-Living Conditions. Diabetes Technology and Therapeutics. 2013;15(8):634-43.
[290] Marling CR, Struble NW, Bunescu RC, Shubrook JH, Schwartz FL. A consensus-perceived glycemic variability metric. Journal of Diabetes Science and Technology. 2013;7 (4):871-79.
[291] Nguyen C, Varney MD, Harrison LC, Morahan G. Definition of high-risk type 1 diabetes HLA-DR and HLA-DQ types using only three single nucleotide polymorphisms. Diabetes. 2013;62(6):2135-40.
[292] Wei Z, Wang K, Qu HQ, Zhang H, Bradfield J, Kim C, et al. From disease association to risk assessment: An optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genetics. 2009;5 (10)(e1000678).
[293] Jensen MH, Mahmoudi Z, Christensen TF, Tarnow L, Seto E, Johansen MD, et al. Evaluation of an Algorithm for Retrospective Hypoglycemia Detection Using Professional Continuous Glucose Monitoring Data. J Diabetes Sci Technol. 2014;8(1):117-22.
[294] Schwartz FL, Shubrook JH, Marling CR. Use of case-based reasoning to enhance intensive management of patients on insulin pump therapy. J Diabetes Sci Technol. 2008;2(4):603-11.
[295] Cordelli E, Maulucci G, De Spirito M, Rizzi A, Pitocco D, Soda P. A decision support system for type 1 diabetes mellitus diagnostics based on dual channel analysis of red blood cell membrane fluidity. Comput Methods Programs Biomed. 2018;162:263-71.
[296] Sampath S, Tkachenko P, Renard E, Pereverzev SV. Glycemic Control Indices and Their Aggregation in the Prediction of Nocturnal Hypoglycemia From Intermittent Blood Glucose Measurements. J Diabetes Sci Technol. 2016;10(6):1245-50.
[297] Georga EI, Protopappas VC, Polyzos D, Fotiadis DI. Evaluation of short-term predictors of glucose concentration in type 1 diabetes combining feature ranking with regression models. Med Biol Eng Comput. 2015;53(12):1305-18.
[298] Ling SH, San PP, Nguyen HT. Non-invasive hypoglycemia monitoring system using extreme learning machine for Type 1 diabetes. ISA Transactions. 2016;64:440-6.
[299] Perez-Gandia C, Garcia-Saez G, Subias D, Rodriguez-Herrero A, Gomez EJ, Rigla M, et al. Decision Support in Diabetes Care: The Challenge of Supporting Patients in Their Daily Living Using a Mobile Glucose Predictor. Journal of Diabetes Science and Technology. 2018;12(2):243-50.
[300] Ceccarelli F, Sciandrone M, Perricone C, Galvan G, Cipriano E, Galligari A, et al. Biomarkers of erosive arthritis in systemic lupus erythematosus: Application of machine learning models. PLoS ONE. 2018;13 (12) (e0207926).
[301] Turner CA, Jacobs AD, Marques CK, Oates JC, Kamen DL, Anderson PE, et al. Word2Vec inversion and traditional text classifiers for phenotyping lupus. BMC medical informatics and decision making. 2017;17(1):126.

[302] Ceccarelli F, Sciandrone M, Perricone C, Galvan G, Morelli F, Vicente LN, et al. Prediction of chronic damage in systemic lupus erythematosus by using machine-learning models. PLoS ONE. 2017;12 (3)(e0174200).
[303] Kan H, Nagar S, Patel J, Wallace DJ, Molta C, Chang DJ. Longitudinal Treatment Patterns and Associated Outcomes in Patients with Newly Diagnosed Systemic Lupus Erythematosus. Clinical Therapeutics. 2016;38(3):610-24.
[304] Wolf BJ, Spainhour JC, Arthur JM, Janech MG, Petri M, Oates JC. Development of Biomarker Models to Predict Outcomes in Lupus Nephritis. Arthritis and Rheumatology. 2016;68(8):1955-63.
[305] Guy RT, Santago P, Langefeld CD. Bootstrap Aggregating of Alternating Decision Trees to Detect Sets of SNPs That Associate With Disease. Genetic Epidemiology. 2012;36(2):99-106.
[306] Tang H, Poynton MR, Hurdle JF, Baird BC, Koford JK, Goldfarb-Rumyantzev AS. Predicting three-year kidney graft survival in recipients with systemic lupus erythematosus. ASAIO journal (American Society for Artificial Internal Organs : 1992). 2011;57(4):300-9.
[307] Armañanzas R, Calvo B, Inza I, López-Hoyos M, Martínez-Taboada V, Ucar E, et al. Microarray Analysis of Autoimmune Diseases by Machine Learning Procedures. IEEE Transactions on Information Technology in Biomedicine. 2009;13(3):341-50.
[308] Huang Z, Shi Y, Cai B, Wang L, Wu Y, Ying B, et al. MALDI-TOF MS combined with magnetic beads for detecting serum protein biomarkers and establishment of boosting decision tree model for diagnosis of systemic lupus erythematosus. Rheumatology. 2009;48(6):626-31.
[309] Murray SG, Avati A, Schmajuk G, Yazdany J. Automated and flexible identification of complex disease: building a model for systemic lupus erythematosus using noisy labeling. J Am Med Inform Assoc. 2018;26(1):61-5.
[310] Reddy BK, Delen D. Predicting hospital readmission for lupus patients: An RNN-LSTM-based deep-learning methodology. Comput Biol Med. 2018;101:199-209.
[311] Tang Y, Zhang W, Zhu M, Zheng L, Xie L, Yao Z, et al. Lupus nephritis pathology prediction with clinical indices. Scientific Reports. 2018;8(1):10231.
[312] Scully M, Anderson B, Lane T, Gasparovic C, Magnotta V, Sibbitt W, et al. An Automated Method for Segmenting White Matter Lesions through Multi-Level Morphometric Feature Classification with Application to Lupus. Front Hum Neurosci. 2010;4:27.
[313] Davis NA, Lareau CA, White BC, Pandey A, Wiley G, Montgomery CG, et al. Encore: Genetic Association Interaction Network centrality pipeline and application to SLE exome data. Genetic Epidemiology. 2013;37(6):614-21.
[314] Wang Y, Li Y, Pu W, Wen K, Shugart YY, Xiong M, et al. Random Bits Forest: a Strong Classifier/Regressor for Big Data. Scientific reports. 2016;6:30086.
[315] George Y, Aldeen M, Garnavi R. Psoriasis image representation using patch-based dictionary learning for erythema severity scoring. Computerized Medical Imaging & Graphics. 2018;66:44-55.
[316] Shrivastava VK, Londhe ND, Sonawane RS, Suri JS. A novel and robust Bayesian approach for segmentation of psoriasis lesions and its risk stratification. Comput Methods Programs Biomed. 2017;150:9-22.
[317] Shrivastava VK, Londhe ND, Sonawane RS, Suri JS. Computer-aided diagnosis of psoriasis skin images with HOS, texture and color features: A first comparative study of its kind. Comput Methods Programs Biomed. 2016;126:98-109.
[318] Shrivastava VK, Londhe ND, Sonawane RS, Suri JS. A novel approach to multiclass psoriasis disease risk stratification: Machine learning paradigm. Biomedical Signal Processing and Control. 2016;28:27-40.
[319] Shrivastava VK, Londhe ND, Sonawane RS, Suri JS. Exploring the color feature power for psoriasis risk stratification and classification: A data mining paradigm. Computers in Biology and Medicine. 2015;65:54-68.
[320] Cowen EW, Liu CW, Steinberg SM, Kang S, Vonderheid EC, Kwak HS, et al. Differentiation of tumour-stage mycosis fungoides, psoriasis vulgaris and normal controls in a pilot study using serum proteomic analysis. British Journal of Dermatology. 2007;157(5):946-53.

[321] Raina A, Hennessy R, Rains M, Allred J, Hirshburg JM, Diven DG, et al. Objective measurement of erythema in psoriasis using digital color photography with color calibration. Skin Res Technol. 2016;22(3):375-80.
[322] Shrivastava VK, Londhe ND, Sonawane RS, Suri JS. Reliable and accurate psoriasis disease classification in dermatology images using comprehensive feature space in machine learning paradigm. Expert Systems with Applications. 2015;42(15/16):6184-95.
[323] Shrivastava VK, Londhe ND, Sonawane RS, Suri JS. Reliability analysis of psoriasis decision support system in principal component analysis framework. Data & Knowledge Engineering. 2016;106:1-17.
[324] Patrick MT, Stuart PE, Raja K, Gudjonsson JE, Tejasvi T, Yang J, et al. Genetic signature to provide robust risk assessment of psoriatic arthritis development in psoriasis patients. Nature Communications. 2018;9 (1) (4178).
[325] Hujoel IA, Murphree DH, Van Dyke CT, Choung RS, Sharma A, Murray JA, et al. Machine Learning in Detection of Undiagnosed Celiac Disease. Clinical Gastroenterology and Hepatology. 2018;16(8):1354-5.e1.
[326] Arasaradnam RP, Westenbrink E, McFarlane MJ, Harbord R, Chambers S, O'Connell N, et al. Differentiating coeliac disease from irritable bowel syndrome by urinary volatile organic compound analysis - A pilot study. PLoS ONE. 2014;9 (10) (e107312).
[327] Tenorio JM, Hummel AD, Cohrs FM, Sdepanian VL, Pisa IT, De Fatima Marin H. Artificial intelligence techniques applied to the development of a decision-support system for diagnosing celiac disease. International Journal of Medical Informatics. 2011;80(11):793-802.
[328] Choung RS, Rostamkolaei SK, Ju JM, Marietta EV, Van Dyke CT, Rajasekaran JJ, et al. Synthetic Neoepitopes of the Transglutaminase-Deamidated Gliadin Complex as Biomarkers for Diagnosing and Monitoring Celiac Disease. Gastroenterology. 2019;156(3):582-91.e1.
[329] Chen W, Huang Y, Boyle B, Lin S. The utility of including pathology reports in improving the computational identification of patients. J Pathol Inform. 2016;7:46.
[330] Ludvigsson JF, Pathak J, Murphy S, Durski M, Kirsch PS, Chute CG, et al. Use of computerized algorithm to identify individuals in need of testing for celiac disease. J Am Med Inform Assoc. 2013;20(e2):e306-10.
[331] Amirkhani A, Mosavi MR, Mohammadi K, Papageorgiou EI. A novel hybrid method based on fuzzy cognitive maps and fuzzy clustering algorithms for grading celiac disease. Neural Computing & Applications. 2018;30(5):1573-88.
[332] Ahmad W, Ahmad A, Lu C, Khoso BA, Huang L. A novel hybrid decision support system for thyroid disease forecasting. Soft Computing - A Fusion of Foundations, Methodologies & Applications. 2018;22(16):5377-83.
[333] Baccour L. Amended fused TOPSIS-VIKOR for classification (ATOVIC) applied to some UCI data sets. Expert Systems with Applications. 2018;99:115-25.
[334] Morejón R, Viana M, Lucena C. An Approach to Generate Software Agents for Health Data Mining. International Journal of Software Engineering & Knowledge Engineering. 2017;27(9/10):1579-89.
[335] Temurtas F. A comparative study on thyroid disease diagnosis using neural networks. Expert Systems with Applications. 2009;36(1):944-9.
[336] Polat K, Şahan S, Güneş S. A novel hybrid method based on artificial immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid disease diagnosis. Expert Systems with Applications. 2007;32(4):1141-7.
[337] Keleş A, Keleş A. ESTDD: Expert system for thyroid diseases diagnosis. Expert Systems with Applications. 2008;34(1):242-6.
[338] Weiss J, Kuusisto F, Boyd K, Liu J, Page D. Machine Learning for Treatment Assignment: Improving Individualized Risk Attribution. Amia 2015;Annual Symposium proceedings. AMIA Symposium. 2015:1306-15.
[339] Singh A, Pandey B. A KLD-LSSVM based computational method applied for feature ranking and classification of primary biliary cirrhosis stages. International Journal of Computational Biology and Drug Design. 2017;10(1):24-38.

[340] Eaton JE, Vesterhus M, McCauley BM, Atkinson EJ, Schlicht EM, Juran BD, et al. Primary Sclerosing Cholangitis Risk Estimate Tool (PREsTo) Predicts Outcomes in PSC: A Derivation & Validation Study Using Machine Learning. Hepatology. 2018.
[341] Iwasawa K, Suda W, Tsunoda T, Oikawa-Kawamoto M, Umetsu S, Takayasu L, et al. Dysbiosis of the salivary microbiota in pediatric-onset primary sclerosing cholangitis and its potential as a biomarker. Scientific Reports. 2018;8(1):5480.
[342] Tsujitani M, Sakon M. Analysis of Survival Data Having Time-Dependent Covariates. IEEE Transactions on Neural Networks. 2009;20(3):389-94.
[343] Zhu H, Zhu C, Mi W, Chen T, Zhao H, Zuo X, et al. Integration of Genome-Wide DNA Methylation and Transcription Uncovered Aberrant Methylation-Regulated Genes and Pathways in the Peripheral Blood Mononuclear Cells of Systemic Sclerosis. International Journal of Rheumatology. 2018;2018 (no pagination)(7342472).
[344] Taroni JN, Martyanov V, Mahoney JM, Whitfield ML. A Functional Genomic Meta-Analysis of Clinical Trials in Systemic Sclerosis: Toward Precision Medicine and Combination Therapy. Journal of Investigative Dermatology. 2017;137(5):1033-41.
[345] Huang H, Fava A, Guhr T, Cimbro R, Rosen A, Boin F, et al. A methodology for exploring biomarker--phenotype associations: application to flow cytometry data and systemic sclerosis clinical manifestations. BMC bioinformatics. 2015;16:293.
[346] Berks M, Tresadern P, Dinsdale G, Murray A, Moore T, Herrick A, et al. An automated system for detecting and measuring nailfold capillaries. Medical image computing and computer-assisted intervention : MICCAI 2014;International Conference on Medical Image Computing and Computer-Assisted Intervention. 17(Pt 1):658-65.
[347] Huang KP, Mullangi S, Guo Y, Qureshi AA. Autoimmune, atopic, and mental health comorbid conditions associated with alopecia areata in the United States.[Erratum appears in JAMA Dermatol. 2014 Jun;150(6):674]. JAMA Dermatology. 2013;149(7):789-94.
[348] Sheth VM, Guo Y, Qureshi AA. Comorbidities associated with vitiligo: A ten-year retrospective study. Dermatology. 2013;227(4):311-5.
[349] Armananzas R, Calvo B, Inza I, Lopez-Hoyos M, Martinez-Taboada V, Ucar E, et al. Microarray analysis of autoimmune diseases by machine learning procedures. IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society. 2009;13(3):341-50.
[350] Gottesman O, Kuivaniemi H, Tromp G, Faucett WA, Li R, Manolio TA, et al. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. Genetics in Medicine. 2013;15(10):761-71.
[351] Nelson CA, Butte AJ, Baranzini SE. Integrating biomedical research and electronic health records to create knowledge-based biologically meaningful machine-readable embeddings. Nature Communications. 2019;10(1):3045.
[352] Raphael BJ, Dobson JR, Oesper L, Vandin F. Identifying driver mutations in sequenced cancer genomes: computational approaches to enable precision medicine. Genome Medicine. 2014;6(1):5.
[353] Zeng Z, Vo AH, Mao C, Clare SE, Khan SA, Luo Y. Cancer classification and pathway discovery using non-negative matrix factorization. Journal of Biomedical Informatics. 2019;96:103247.
[354] Zhang X, Guan N, Jia Z, Qiu X, Luo Z. Semi-Supervised Projective Non-Negative Matrix Factorization for Cancer Classification. PLOS ONE. 2015;10(9):e0138814.
[355] Huang J, Ling CX. Using AUC and accuracy in evaluating learning algorithms. Ieee T Knowl Data En. 2005;17(3):299-310.
[356] Pengelly RJ, Gibson J, Andreoletti G, Collins A, Mattocks CJ, Ennis S. A SNP profiling panel for sample tracking in whole-exome sequencing studies. Genome Medicine. 2013;5(9):89.
[357] Mossotto E, Ashton JJ, O'Gorman L, Pengelly RJ, Beattie RM, MacArthur BD, et al. GenePy - a score for estimating gene pathogenicity in individuals using next-generation sequencing data. BMC Bioinformatics. 2019;20(1):254.

[358] Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv. 2013;1303.
[359] Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current protocols in bioinformatics. 2013;43(1110):11.0.1-.0.33.
[360] Poplin R, Ruano-Rubio V, DePristo MA, Fennell TJ, Carneiro MO, Van der Auwera GA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018:201178.
[361] Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research. 2010;38(16):e164.
[362] McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics. 2010;26(16):2069-70.
[363] Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Medicine. 2021;13(1):31.
[364] Cheng J, Nguyen TYD, Cygan KJ, Çelik MH, Fairbrother WG, Avsec ž, et al. MMSplice: modular modeling improves the predictions of genetic variant effects on splicing. Genome Biology. 2019;20(1):48.
[365] Jaganathan K, Kyriazopoulou Panagiotopoulou S, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting Splicing from Primary Sequence with Deep Learning. Cell. 2019;176(3):535-48.e24.
[366] Salem M, Ammitzboell M, Nys K, Seidelin JB, Nielsen OH. ATG16L1: A multifunctional susceptibility factor in Crohn disease. Autophagy. 2015;11(4):585-94.
[367] Glas J, Konrad A, Schmechel S, Dambacher J, Seiderer J, Schroff F, et al. The ATG16L1 Gene Variants rs2241879 and rs2241880 (T300A) Are Strongly Associated With Susceptibility to Crohn's Disease in the German Population. Official journal of the American College of Gastroenterology | ACG. 2008;103(3):682-91.
[368] Ignatieva E, Levitsky V, Yudin N, Moshkin M, Kolchanov N. Genetic basis of olfactory cognition: extremely high level of DNA sequence polymorphism in promoter regions of the human olfactory receptor genes revealed using the 1000 Genomes Project dataset. Frontiers in Psychology. 2014;5.
[369] Levine A, Koletzko S, Turner D, Escher JC, Cucchiara S, de Ridder L, et al. ESPGHAN revised porto criteria for the diagnosis of inflammatory bowel disease in children and adolescents. J Pediatr Gastroenterol Nutr. 2014;58(6):795-806.
[370] Lamb CA, Kennedy NA, Raine T, Hendy PA, Smith PJ, Limdi JK, et al. British Society of Gastroenterology consensus guidelines on the management of inflammatory bowel disease in adults. Gut. 2019;68(Suppl 3):s1-s106.
[371] Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. Nucleic acids research. 1988;16(3):1215-.
[372] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.
[373] Picard Tools. 1.97 ed. http://broadinstitute.github.io/picard/: Broad Institute.
[374] Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Current protocols in bioinformatics. 2013;43:11.0.1-33.
[375] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic acids research. 2016;44(D1):D733-45.
[376] Liu X, Jian X, Boerwinkle E. dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Human Mutation. 2011;32(8):894-9.
[377] Jun G, Flickinger M, Hetrick KN, Romm JM, Doheny KF, Abecasis GR, et al. Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. Am J Hum Genet. 2012;91(5):839-48.
[378] Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841-2.

[379] Carson AR, Smith EN, Matsui H, Brækkan SK, Jepsen K, Hansen J-B, et al. Effective filtering strategies to improve data quality from population-based whole exome sequencing studies. BMC bioinformatics. 2014;15:125-.
[380] Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156-8.
[381] Hunt SE, McLaren W, Gil L, Thormann A, Schuilenburg H, Sheppard D, et al. Ensembl variation resources. Database. 2018;2018.
[382] UK HDR. Gut Reaction Health Data Research Hub 2022 [Available from: https://gut-reaction.org/ibd-data-gut-reaction/.
[383] Ryan P, Valentin R-R, Mark AD, Tim JF, Mauricio OC, Geraldine AVdA, et al. Scaling accurate genetic variant discovery to tens of thousands of samples. bioRxiv. 2018:201178.
[384] Geraldine A. Van der Auwera BDOC. Genomics in the Cloud: Using Docker, GATK, and WDL in Terra (1st Edition): O'Reilly Media.; 2020.
[385] Dale RK, Pedersen BS, Quinlan AR. Pybedtools: a flexible Python library for manipulating genomic datasets and annotations. Bioinformatics. 2011;27(24):3423-4.
[386] Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization of intersecting sets and their properties. Bioinformatics. 2017;33(18):2938-40.
[387] Ashton JJ, Boukas K, Stafford IS, Cheng G, Haggarty R, Coelho TAF, et al. Deleterious Genetic Variation Across the NOD Signaling Pathway Is Associated With Reduced NFKB Signaling Transcription and Upregulation of Alternative Inflammatory Transcripts in Pediatric Inflammatory Bowel Disease. Inflammatory Bowel Diseases. 2022;28(6):912-22.
[388] Ashton JJ, Cheng G, Stafford IS, Kellermann M, Seaby EG, Cummings, J R Fraser, et al. Prediction of Crohn's Disease Stricturing Phenotype Using a NOD2-derived Genomic Biomarker. Inflammatory Bowel Diseases. 2022;29(4):511-21.
[389] Mossotto E, Boberska J, Ashton JJ, Stafford IS, Cheng G, Baker J, et al. Evidence of a genetically driven metabolomic signature in actively inflamed Crohn's disease. Scientific Reports. 2022;12(1):14101.
[390] Coelho T, Sonnenberg-Riethmacher E, Gao Y, Mossotto E, Khojanazarov A, Griffin A, et al. Expression profile of the matricellular protein periostin in paediatric inflammatory bowel disease. Scientific Reports. 2021;11(1):6194.
[391] Coelho T, Mossotto E, Gao Y, Haggarty R, Ashton JJ, Batra A, et al. Immunological Profiling of Paediatric Inflammatory Bowel Disease Using Unsupervised Machine Learning. J Pediatr Gastroenterol Nutr. 2020;70(6):833-40.
[392] Liu Z, Ren Z, Zhang J, Chuang C-C, Kandaswamy E, Zhou T, et al. Role of ROS and Nutritional Antioxidants in Human Diseases. Front Physiol. 2018;9:477-.
[393] Mateen S, Moin S, Khan AQ, Zafar A, Fatima N. Increased Reactive Oxygen Species Formation and Oxidative Stress in Rheumatoid Arthritis. PLOS ONE. 2016;11(4):e0152925.
[394] Lipińska J, Lipińska S, Stańczyk J, Sarniak A, Przymińska vel Prymont A, Kasielski M, et al. Reactive oxygen species and serum antioxidant defense in juvenile idiopathic arthritis. Clin Rheumatol. 2015;34(3):451-6.
[395] Castro-Correia C, Maia ML, Norberto S, Costa-Santos C, Barroso MF, Carvalho A, et al. Can Antioxidative Status Be Involved in Type 1 Diabetes? J Clin Med Res. 2017;9(12):998-1001.
[396] Juybari KB, Ebrahimi G, Momeni Moghaddam MA, Asadikaram G, Torkzadeh-Mahani M, Akbari M, et al. Evaluation of serum arsenic and its effects on antioxidant alterations in relapsing-remitting multiple sclerosis patients. Multiple Sclerosis and Related Disorders. 2018;19:79-84.
[397] Picca A, Riezzo G, Lezza AMS, Clemente C, Pesce V, Orlando A, et al. Mitochondria and redox balance in coeliac disease: A case-control study. European journal of clinical investigation. 2018;48(2).
[398] Ademoğlu E, Özbey N, Erbil Y, Tanrikulu S, Barbaros U, Yanik BT, et al. Determination of oxidative stress in thyroid tissue and plasma of patients with Graves' disease. European Journal of Internal Medicine. 2006;17(8):545-50.

[399] van de Veerdonk FL, Dinarello CA. Deficient autophagy unravels the ROS paradox in chronic granulomatous disease. Autophagy. 2014;10(6):1141-2.
[400] Szczeklik K, Krzyściak W, Cibor D, Domagała-Rodacka R, Pytko-Polończyk J, Mach T, et al. Markers of lipid peroxidation and antioxidant status in the serum and saliva of patients with active Crohn disease. Polish archives of internal medicine. 2018;128(6):362-70.
[401] Szczeklik K, Owczarek D, Cibor D, Cześnikiewicz-Guzik M, Krzyściak P, Krawczyk A, et al. Relative homogeneity of oral bacterial oral in Crohn's disease compared to ulcerative colitis and its connections with antioxidant defense - preliminary report. Folia medica Cracoviensia. 2019;59(1):15-35.
[402] Luceri C, Bigagli E, Agostiniani S, Giudici F, Zambonin D, Scaringi S, et al. Analysis of Oxidative Stress-Related Markers in Crohn's Disease Patients at Surgery and Correlations with Clinical Findings. Antioxidants (Basel, Switzerland). 2019;8(9).
[403] van Langenberg DR, Della Gatta P, Warmington SA, Kidgell DJ, Gibson PR, Russell AP. Objectively measured muscle fatigue in Crohn's disease: Correlation with self-reported fatigue and associated factors for clinical application. Journal of Crohn's and Colitis. 2014;8(2):137-46.
[404] Maor I, Rainis T, Lanir A, Lavy A. Oxidative stress, inflammation and neutrophil superoxide release in patients with Crohn's disease: distinction between active and non-active disease. Digestive diseases and sciences. 2008;53(8):2208-14.
[405] Jahanshahi G, Motavasel V, Rezaie A, Hashtroudi AA, Daryani NE, Abdollahi M. Alterations in antioxidant power and levels of epidermal growth factor and nitric oxide in saliva of patients with inflammatory bowel diseases. Digestive diseases and sciences. 2004;49(11-12):1752-7.
[406] Norouzinia M, Chaleshi V, Alizadeh AHM, Zali MR. Biomarkers in inflammatory bowel diseases: insight into diagnosis, prognosis and treatment. Gastroenterol Hepatol Bed Bench. 2017;10(3):155-67.
[407] Pathirana WGW, Chubb SP, Gillett MJ, Vasikaran SD. Faecal Calprotectin. Clin Biochem Rev. 2018;39(3):77-90.
[408] Benzie IF, Strain JJ. The ferric reducing ability of plasma (FRAP) as a measure of "antioxidant power": the FRAP assay. Analytical biochemistry. 1996;239(1):70-6.
[409] Espin S, Sanchez Virosta P, García-Fernández A, Eeva T. A microplate adaptation of the thiobarbituric acid reactive substances assay to determine lipid peroxidation fluorometrically in small sample volumes. Revista de Toxicologia. 2017;34:In press.
[410] Panday A, Sahoo MK, Osorio D, Batra S. NADPH oxidases: an overview from structure to innate immunity-associated pathologies. Cellular & molecular immunology. 2015;12(1):5-23.
[411] Kuhn M. Building Predictive Models in R Using the caret Package. Journal of Statistical Software. 2008;028(i05).
[412] Schwerd T, Bryant RV, Pandey S, Capitani M, Meran L, Cazier JB, et al. NOX1 loss-of-function genetic variants in patients with inflammatory bowel disease. Mucosal Immunology. 2017;11:562.
[413] van Beelen Granlund A, Østvik AE, Brenna Ø, Torp SH, Gustafsson BI, Sandvik AK. REG gene expression in inflamed and healthy colon mucosa explored by in situ hybridisation. Cell Tissue Res. 2013;352(3):639-46.
[414] Poss KD, Tonegawa S. Reduced stress defense in heme oxygenase 1-deficient cells. Proceedings of the National Academy of Sciences of the United States of America. 1997;94(20):10925-30.
[415] Economou M, Trikalinos TA, Loizou KT, Tsianos EV, Ioannidis JP. Differential effects of NOD2 variants on Crohn's disease risk and phenotype in diverse populations: a metaanalysis. The American journal of gastroenterology. 2004;99(12):2393-404.
[416] Muise AM, Xu W, Guo C-H, Walters TD, Wolters VM, Fattouh R, et al. NADPH oxidase complex and IBD candidate gene studies: identification of a rare variant in NCF2 that results in reduced binding to RAC2. Gut. 2012;61(7):1028-35.
[417] KUMAR D, PANDEY RK, AGRAWAL D, AGRAWAL D. An estimation and evaluation of total antioxidant capacity of saliva in children with severe early childhood caries. International Journal of Paediatric Dentistry. 2011;21(6):459-64.

[418] Lipinski S, Petersen BS, Barann M, Piecyk A, Tran F, Mayr G, et al. Missense variants in NOX1 and p22phox in a case of very-early-onset inflammatory bowel disease are functionally linked to NOD2. Cold Spring Harbor molecular case studies. 2019;5(1).

[419] Caruso R, Mathes T, Martens EC, Kamada N, Nusrat A, Inohara N, et al. A specific gene-microbe interaction drives the development of Crohn's disease–like colitis in mice. Science Immunology. 2019;4(34):eaaw4341.

[420] Zaharie R, Tantau A, Zaharie F, Tantau M, Gheorghe L, Gheorghe C, et al. Diagnostic Delay in Romanian Patients with Inflammatory Bowel Disease: Risk Factors and Impact on the Disease Course and Need for Surgery. Journal of Crohn's & colitis. 2016;10(3):306-14.

[421] Moon CM, Jung SA, Kim SE, Song HJ, Jung Y, Ye BD, et al. Clinical Factors and Disease Course Related to Diagnostic Delay in Korean Crohn's Disease Patients: Results from the CONNECT Study. PLoS One. 2015;10(12):e0144390.

[422] Blackwell J, Saxena S, Jayasooriya N, Bottle A, Petersen I, Hotopf M, et al. Prevalence and Duration of Gastrointestinal Symptoms Before Diagnosis of Inflammatory Bowel Disease and Predictors of Timely Specialist Review: A Population-Based Study. Journal of Crohn's and Colitis. 2020;15(2):203-11.

[423] National Institute for Health and Care Excellence. Inflammatory bowel disease Quality Standard [QS81] 2015 [Available from: https://www.nice.org.uk/guidance/qs81/chapter/quality-statement-1-specialist-assessment.

[424] Crohn's & Colitis UK. Tests and Investigations 2022 [Available from: https://crohnsandcolitis.org.uk/info-support/information-about-crohns-and-colitis/all-information-about-crohns-and-colitis/healthcare/tests-and-investigations.

[425] Walker GJ, Moore L, Heerasing N, Hendy P, Perry MH, McDonald TJ, et al. Faecal calprotectin effectively excludes inflammatory bowel disease in 789 symptomatic young adults with/without alarm symptoms: a prospective UK primary care cohort study. Alimentary Pharmacology & Therapeutics. 2018;47(8):1103-16.

[426] National Institute of Diabetes and Digestive and Kidney Diseases. Diagnosis of Ulcerative Colitis 2020 [Available from: https://www.niddk.nih.gov/health-information/digestive-diseases/ulcerative-colitis/diagnosis.

[427] National Institute of Diabetes and Digestive and Kidney Diseases. Diagnosis of Crohn's Disease 2017 [Available from: https://www.niddk.nih.gov/health-information/digestive-diseases/crohns-disease/diagnosis.

[428] Pillai N, Lupatsch JE, Dusheiko M, Schwenkglenks M, Maillard M, Sutherland CS, et al. Evaluating the Cost-Effectiveness of Early Compared with Late or No Biologic Treatment to Manage Crohn's Disease using Real-World Data. Journal of Crohn's and Colitis. 2019;14(4):490-500.

[429] Tsui JJ, Huynh HQ. Is top-down therapy a more effective alternative to conventional step-up therapy for Crohn's disease? Ann Gastroenterol. 2018;31(4):413-24.

[430] Mohan HM, Coffey JC. Surgical treatment of intestinal stricture in inflammatory bowel disease. Journal of Digestive Diseases. 2020;21(6):355-9.

[431] Bessissow T, Reinglas J, Aruljothy A, Lakatos PL, Van Assche G. Endoscopic management of Crohn's strictures. World journal of gastroenterology. 2018;24(17):1859-67.

[432] Waljee AK, Wallace BI, Cohen-Mekelburg S, Liu Y, Liu B, Sauder K, et al. Development and Validation of Machine Learning Models in Prediction of Remission in Patients With Moderate to Severe Crohn Disease. JAMA network open. 2019;2(5):e193721.

[433] Dong Y, Xu L, Fan Y, Xiang P, Gao X, Chen Y, et al. A novel surgical predictive model for Chinese Crohn's disease patients. Medicine. 2019;98(46):e17510.

[434] Menti E, Lanera C, Lorenzoni G, Giachino DF, Marchi M, Gregori D, et al. Bayesian Machine Learning Techniques for revealing complex interactions among genetic and clinical factors in association with extra-intestinal Manifestations in IBD patients. AMIA Annu Symp Proc. 2016;2016:884-93.

[435] Bottigliengo D, Berchialla P, Lanera C, Azzolina D, Lorenzoni G, Martinato M, et al. The Role of Genetic Factors in Characterizing Extra-Intestinal Manifestations in Crohn's Disease Patients: Are Bayesian Machine Learning Methods Improving Outcome Predictions? Journal of clinical medicine. 2019;8(6).
[436] Wei Z, Wang W, Bradfield J, Li J, Cardinale C, Frackelton E, et al. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. Am J Hum Genet. 2013;92(6):1008-12.
[437] Romagnoni A, Jégou S, Van Steen K, Wainrib G, Hugot JP. Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. Sci Rep. 2019;9(1):10351.
[438] Taylor KM, Hanscombe KB, Prescott NJ, Iniesta R, Traylor M, Taylor NS, et al. Genetic and Inflammatory Biomarkers Classify Small Intestine Inflammation in Asymptomatic First-degree Relatives of Patients With Crohn's Disease. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association. 2020;18(4):908-16.e13.
[439] Giollo M, Jones DT, Carraro M, Leonardi E, Ferrari C, Tosatto SCE. Crohn disease risk prediction-Best practices and pitfalls with exome data. Hum Mutat. 2017;38(9):1193-200.
[440] Wang Y, Miller M, Astrakhan Y, Petersen BS, Schreiber S, Franke A, et al. Identifying Crohn's disease signal from variome analysis. Genome Med. 2019;11(1):59.
[441] Raimondi D, Simm J, Arany A, Fariselli P, Cleynen I, Moreau Y. An interpretable low-complexity machine learning framework for robust exome-based in-silico diagnosis of Crohn's disease patients. NAR Genom Bioinform. 2020;2(1):lqaa011.
[442] Fuentes Fajardo KV, Adams D, Program NCS, Mason CE, Sincan M, Tifft C, et al. Detecting false-positive signals in exome sequencing. Human Mutation. 2012;33(4):609-13.
[443] Pedersen BS, Quinlan AR. Who's Who? Detecting and Resolving Sample Anomalies in Human DNA Sequencing Studies with *Peddy*. The American Journal of Human Genetics. 2017;100(3):406-13.
[444] van der Velde KJ, de Boer EN, van Diemen CC, Sikkema-Raddatz B, Abbott KM, Knopperts A, et al. GAVIN: Gene-Aware Variant INterpretation for medical sequencing. Genome Biology. 2017;18(1):6.
[445] Belot A, Rice GI, Omarjee SO, Rouchon Q, Smith EMD, Moreews M, et al. Contribution of rare and predicted pathogenic gene variants to childhood-onset lupus: a large, genetic panel analysis of British and French cohorts. The Lancet Rheumatology. 2020;2(2):e99-e109.
[446] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825-30.
[447] Committee HGN. 2021 [Available from: https://www.genenames.org/tools/multi-symbol-checker/.
[448] Raudvere U, Kolberg L, Kuzmin I, Arak T, Adler P, Peterson H, et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). Nucleic acids research. 2019;47(W1):W191-W8.
[449] GeneCards - the human gene database 2022 [cited 2022 September 10]. Available from: www.genecards.org.
[450] Stelzer G, Rosen N, Plaschkes I, Zimmerman S, Twik M, Fishilevich S, et al. The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. Current protocols in bioinformatics. 2016;54:1.30.1-1..3.
[451] National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information [Available from: https://www.ncbi.nlm.nih.gov/gene.
[452] Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic acids research. 2022;50(D1):D20-d6.
[453] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems; Long Beach, California, USA: Curran Associates Inc.; 2017. p. 4768–77.
[454] Chen EY, Tan CM, Kou Y, Duan Q, Wang Z, Meirelles GV, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. BMC Bioinformatics. 2013;14(1):128.

[455] Kanehisa M, Furumichi M, Sato Y, Ishiguro-Watanabe M, Tanabe M. KEGG: integrating viruses and cellular organisms. Nucleic acids research. 2020;49(D1):D545-D51.
[456] Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic acids research. 2020;49(D1):D884-D91.
[457] Panés J, Vermeire S. JAK Inhibitors: Back to Small Molecules for the Treatment of IBD. Journal of Crohn's and Colitis. 2020;14(Supplement_2):S711-S2.
[458] Andreoletti G, Shakhnovich V, Christenson K, Coelho T, Haggarty R, Afzal NA, et al. Exome Analysis of Rare and Common Variants within the NOD Signaling Pathway. Sci Rep. 2017;7:46454.
[459] Ramos PS, Criswell LA, Moser KL, Comeau ME, Williams AH, Pajewski NM, et al. A Comprehensive Analysis of Shared Loci between Systemic Lupus Erythematosus (SLE) and Sixteen Autoimmune Diseases Reveals Limited Genetic Overlap. PLOS Genetics. 2011;7(12):e1002406.
[460] Yuan Q, Li Y, Li J, Bian X, Long F, Duan R, et al. WDFY4 Is Involved in Symptoms of Systemic Lupus Erythematosus by Modulating B Cell Fate via Noncanonical Autophagy. The Journal of Immunology. 2018;201(9):2570.
[461] Riedl S, Tandara A, Reinshagen M, Hinz U, Faissner A, Bodenmüller H, et al. Serum tenascin-C is an indicator of inflammatory bowel disease activity. Int J Colorectal Dis. 2001;16(5):285-91.
[462] Brant SR, Okou DT, Simpson CL, Cutler DJ, Haritunians T, Bradfield JP, et al. Genome-Wide Association Study Identifies African-Specific Susceptibility Loci in African Americans With Inflammatory Bowel Disease. Gastroenterology. 2017;152(1):206-17.e2.
[463] Vaeth M, Feske S. NFAT control of immune function: New Frontiers for an Abiding Trooper. F1000Research. 2018;7:260.
[464] Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, et al. STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic acids research. 2015;43(Database issue):D447-52.
[465] Fasseu M, Tréton X, Guichard C, Pedruzzi E, Cazals-Hatem D, Richard C, et al. Identification of restricted subsets of mature microRNA abnormally expressed in inactive colonic mucosa of patients with inflammatory bowel disease. PloS one. 2010;5(10):e13160.
[466] Strisciuglio C, Cenni S, Giugliano FP, Miele E, Cirillo G, Martinelli M, et al. The Role of Inflammation on Vitamin D Levels in a Cohort of Pediatric Patients With Inflammatory Bowel Disease. J Pediatr Gastroenterol Nutr. 2018;67(4):501-6.
[467] Ghaly S, Murray K, Baird A, Martin K, Prosser R, Mill J, et al. High Vitamin D–Binding Protein Concentration, Low Albumin, and Mode of Remission Predict Relapse in Crohn's Disease. Inflammatory Bowel Diseases. 2016;22(10):2456-64.
[468] Boccarelli A, Del Buono N, Esposito F. Colorectal cancer in Crohn's disease evaluated with genes belonging to fibroblasts of the intestinal mucosa selected by NMF. Pathology - Research and Practice. 2022;229:153728.
[469] Broom OJ, Widjaya B, Troelsen J, Olsen J, Nielsen OH. Mitogen activated protein kinases: a role in inflammatory bowel disease? Clinical and Experimental Immunology. 2009;158(3):272-80.
[470] National Library of Medicine. MAPT microtubule associated protein tau [Homo sapiens (human)]

2023 [Available from: https://www.ncbi.nlm.nih.gov/gene/4137#summary.
[471] National Library of Medicine. APOL5 apolipoprotein L5 [Homo sapiens (human)] 2023 [Available from: https://www.ncbi.nlm.nih.gov/gene/80831#summary.
[472] National Library of Medicine. PRKRA protein activator of interferon induced protein kinase EIF2AK2 [Homo sapiens (human)] 2023 [Available from: https://www.ncbi.nlm.nih.gov/gene/8575#summary.
[473] Stenke E, Aviello G, Singh A, Martin S, Winter D, Sweeney B, et al. NADPH oxidase 4 is protective and not fibrogenic in intestinal inflammation. Redox Biology. 2020;37:101752.
[474] Imai J, Kitamoto S, Sugihara K, Nagao-Kitamoto H, Hayashi A, Morhardt TL, et al. Flagellin-mediated activation of IL-33-ST2 signaling by a pathobiont promotes intestinal fibrosis. Mucosal Immunology. 2019;12(3):632-43.

[475] Kerur B, Machan JT, Shapiro JM, Cerezo CS, Markowitz J, Mack DR, et al. Biologics Delay Progression of Crohn's Disease, but Not Early Surgery, in Children. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association. 2018;16(9):1467-73.
[476] Ihara S, Hirata Y, Koike K. TGF-β in inflammatory bowel disease: a key regulator of immune cells, epithelium, and the intestinal microbiota. Journal of Gastroenterology. 2017;52(7):777-87.
[477] Li C, Flynn RS, Grider JR, Murthy KS, Kellum JM, Akbari H, et al. Increased Activation of Latent TGF-β1 by αVβ3 in Human Crohn's Disease and Fibrosis in TNBS Colitis Can Be Prevented by Cilengitide. Inflammatory Bowel Diseases. 2013;19(13):2829-39.
[478] Li C, Iness A, Yoon J, Grider JR, Murthy KS, Kellum JM, et al. Noncanonical STAT3 Activation Regulates Excess TGF-β1 and Collagen I Expression in Muscle of Stricturing Crohn's Disease. The Journal of Immunology. 2015;194(7):3422-31.
[479] Goel MK, Khanna P, Kishore J. Understanding survival analysis: Kaplan-Meier estimate. Int J Ayurveda Res. 2010;1(4):274-8.
[480] HARRELL Jr. FE, LEE KL, MARK DB. MULTIVARIABLE PROGNOSTIC MODELS: ISSUES IN DEVELOPING MODELS, EVALUATING ASSUMPTIONS AND ADEQUACY, AND MEASURING AND REDUCING ERRORS. Statistics in Medicine. 1996;15(4):361-87.
[481] Kim S, Kim K, Choe J, Lee I, Kang J. Improved survival analysis by learning shared genomic information from pan-cancer data. Bioinformatics. 2020;36(Suppl_1):i389-i98.
[482] Sanz H, Reverter F, Valim C. Enhancing SVM for survival data using local invariances and weighting. BMC Bioinformatics. 2020;21(1):193.
[483] Ching T, Zhu X, Garmire LX. Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data. PLOS Computational Biology. 2018;14(4):e1006076.
[484] Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. The Annals of Applied Statistics. 2008;2(3):841-60, 20.
[485] Ungaro RC, Hu L, Ji J, Nayar S, Kugathasan S, Denson LA, et al. Machine learning identifies novel blood protein predictors of penetrating and stricturing complications in newly diagnosed paediatric Crohn's disease. Aliment Pharmacol Ther. 2021;53(2):281-90.
[486] Satopaa V, Albrecht J, Irwin D, Raghavan B, editors. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. 2011 31st International Conference on Distributed Computing Systems Workshops; 2011 20-24 June 2011.
[487] Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLOS Medicine. 2015;12(3):e1001779.
[488] Wang J, Chen N, Guo J, Xu X, Liu L, Yi Z. SurvNet: A Novel Deep Neural Network for Lung Cancer Survival Analysis With Missing Values. Front Oncol. 2020;10:588990.
[489] Lee C, Zame W, Yoon J, van der Schaar M. DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks. Proceedings of the AAAI Conference on Artificial Intelligence. 2018;32(1).
[490] Stafford IS, Kellermann M, Mossotto E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. npj Digital Medicine. 2020;3(1):30.
[491] England N. NHS begins new search for AI tools to save lives and improve care 2020 [updated 3 November 2020. Available from: https://transform.england.nhs.uk/news/nhs-begins-new-search-ai-tools-save-lives-and-improve-care/.
[492] Nguyen NH, Picetti D, Dulai PS, Jairath V, Sandborn WJ, Ohno-Machado L, et al. Machine Learning-based Prediction Models for Diagnosis and Prognosis in Inflammatory Bowel Diseases: A Systematic Review. Journal of Crohn's and Colitis. 2021;16(3):398-413.
[493] Tontini GE, Rimondi A, Vernero M, Neumann H, Vecchi M, Bezzio C, et al. Artificial intelligence in gastrointestinal endoscopy for inflammatory bowel disease: a systematic review and new horizons. Therapeutic advances in gastroenterology. 2021;14:17562848211017730-.
[494] Lima CA, Lyra AC, Rocha R, Santana GO. Risk factors for osteoporosis in inflammatory bowel disease patients. World J Gastrointest Pathophysiol. 2015;6(4):210-8.

[495] Liu T, Han L, Tilley M, Afzelius L, Maciejewski M, Jelinsky S, et al. Distinct clinical phenotypes for Crohn's disease derived from patient surveys. BMC Gastroenterol. 2021;21(1):160.
[496] Kieft K, Zhou Z, Anantharaman K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. Microbiome. 2020;8(1):90.
[497] Lerrigo R, Coffey JTR, Kravitz JL, Jadhav P, Nikfarjam A, Shah NH, et al. The Emotional Toll of Inflammatory Bowel Disease: Using Machine Learning to Analyze Online Community Forum Discourse. Crohn's & Colitis 360. 2019;1(2).
[498] Clooney AG, Eckenberger J, Laserna-Mendieta E, Sexton KA, Bernstein MT, Vagianos K, et al. Ranking microbiome variance in inflammatory bowel disease: a large longitudinal intercontinental study. Gut. 2021;70(3):499-510.
[499] Dhaliwal J, Erdman L, Drysdale E, Rinawi F, Muir J, Walters TD, et al. Accurate Classification of Pediatric Colonic Inflammatory Bowel Disease Subtype Using a Random Forest Machine Learning Classifier. J Pediatr Gastroenterol Nutr. 2021;72(2):262-9.
[500] Le V, Quinn TP, Tran T, Venkatesh S. Deep in the Bowel: Highly Interpretable Neural Encoder-Decoder Networks Predict Gut Metabolites from Gut Microbiome. BMC Genomics. 2020;21(Suppl 4):256.
[501] Niehaus KE, Uhlig HH, Clifton DA. Phenotypic characterisation of Crohn's disease severity. Annu Int Conf IEEE Eng Med Biol Soc. 2015;2015:7023-6.
[502] Biernacka KB, Barańska D, Matera K, Podgórski M, Czkwianianc E, Szabelska-Zakrzewska K, et al. The value of magnetic resonance enterography in diagnostic difficulties associated with Crohn's disease. Pol J Radiol. 2021;86:e143-e50.
[503] Volkova A, Ruggles KV. Predictive Metagenomic Analysis of Autoimmune Disease Identifies Robust Autoimmunity and Disease Specific Microbial Signatures. Front Microbiol. 2021;12:621310.
[504] Nuzzo A, Saha S, Berg E, Jayawickreme C, Tocker J, Brown JR. Expanding the drug discovery space with predicted metabolite-target interactions. Commun Biol. 2021;4(1):288.
[505] Xu C, Zhou M, Xie Z, Li M, Zhu X, Zhu H. LightCUD: a program for diagnosing IBD based on human gut microbiome data. BioData Min. 2021;14(1):2.
[506] Manandhar I, Alimadadi A, Aryal S, Munroe PB, Joe B, Cheng X. Gut microbiome-based supervised machine learning for clinical diagnosis of inflammatory bowel diseases. Am J Physiol Gastrointest Liver Physiol. 2021.
[507] Khorasani HM, Usefi H, Peña-Castillo L. Detecting ulcerative colitis from colon samples using efficient feature selection and machine learning. Sci Rep. 2020;10(1):13744.
[508] Jiang P, Lai S, Wu S, Zhao XM, Chen WH. Host DNA contents in fecal metagenomics as a biomarker for intestinal diseases and effective treatment. BMC Genomics. 2020;21(1):348.
[509] Sarrabayrouse G, Elias A, Yáñez F, Mayorga L, Varela E, Bartoli C, et al. Fungal and Bacterial Loads: Noninvasive Inflammatory Bowel Disease Biomarkers for the Clinical Setting. mSystems. 2021;6(2).
[510] Iablokov SN, Klimenko NS, Efimova DA, Shashkova T, Novichkov PS, Rodionov DA, et al. Metabolic Phenotypes as Potential Biomarkers for Linking Gut Microbiome With Inflammatory Bowel Diseases. Front Mol Biosci. 2020;7:603740.
[511] Douglas GM, Hansen R, Jones CMA, Dunn KA, Comeau AM, Bielawski JP, et al. Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. Microbiome. 2018;6(1):13.
[512] Hübenthal M, Hemmrich-Stanisak G, Degenhardt F, Szymczak S, Du Z, Elsharawy A, et al. Sparse Modeling Reveals miRNA Signatures for Diagnostics of Inflammatory Bowel Disease. PLoS One. 2015;10(10):e0140155.
[513] Cui H, Zhang X. Alignment-free supervised classification of metagenomes by recursive SVM. BMC Genomics. 2013;14:641.

[514] Waljee AK, Cohen-Mekelburg S, Liu Y, Liu B, Zhu J, Higgins PDR. Assessing Clinical Disease Recurrence Using Laboratory Data in Surgically Resected Patients From the TOPPIC Trial. Crohn's & Colitis 360. 2020;2(4).
[515] Stidham RW, Liu Y, Enchakalody B, Van T, Krishnamurthy V, Su GL, et al. The Use of Readily Available Longitudinal Data to Predict the Likelihood of Surgery in Crohn Disease. Inflamm Bowel Dis. 2021.
[516] Udristoiu AL, Stefanescu D, Gruionu G, Gruionu LG, Iacob AV, Karstensen JG, et al. Deep Learning Algorithm for the Confirmation of Mucosal Healing in Crohn's Disease, Based on Confocal Laser Endomicroscopy Images. J Gastrointestin Liver Dis. 2021;30(1):59-65.
[517] Sakurai T, Nishiyama H, Sakai K, De Velasco MA, Nagai T, Komeda Y, et al. Mucosal microbiota and gene expression are associated with long-term remission after discontinuation of adalimumab in ulcerative colitis. Sci Rep. 2020;10(1):19186.
[518] Kang EA, Jang J, Choi CH, Kang SB, Bang KB, Kim TO, et al. Development of a Clinical and Genetic Prediction Model for Early Intestinal Resection in Patients with Crohn's Disease: Results from the IMPACT Study. Journal of clinical medicine. 2021;10(4).
[519] Sofo L, Caprino P, Schena CA, Sacchetti F, Potenza AE, Ciociola A. New perspectives in the prediction of postoperative complications for high-risk ulcerative colitis patients: machine learning preliminary approach. Eur Rev Med Pharmacol Sci. 2020;24(24):12781-7.
[520] Shivaji UN, Bazarova A, Critchlow T, Smith SCL, Nardone OM, Love M, et al. Clinical outcomes, predictors of prognosis and health economics consequences in IBD patients after discontinuation of the first biological therapy. Therap Adv Gastroenterol. 2020;13:1756284820981216.
[521] Taylor H, Serrano-Contreras JI, McDonald JAK, Epstein J, Fell JM, Seoane RC, et al. Multiomic features associated with mucosal healing and inflammation in paediatric Crohn's disease. Aliment Pharmacol Ther. 2020;52(9):1491-502.
[522] Choi YI, Park SJ, Chung JW, Kim KO, Cho JH, Kim YJ, et al. Development of Machine Learning Model to Predict the 5-Year Risk of Starting Biologic Agents in Patients with Inflammatory Bowel Disease (IBD): K-CDM Network Study. Journal of clinical medicine. 2020;9(11).
[523] Ghoshal UC, Rai S, Kulkarni A, Gupta A. Prediction of outcome of treatment of acute severe ulcerative colitis using principal component analysis and artificial intelligence. JGH Open. 2020;4(5):889-97.
[524] Jones CMA, Connors J, Dunn KA, Bielawski JP, Comeau AM, Langille MGI, et al. Bacterial Taxa and Functions Are Predictive of Sustained Remission Following Exclusive Enteral Nutrition in Pediatric Crohn's Disease. Inflamm Bowel Dis. 2020;26(7):1026-37.
[525] Braun T, Di Segni A, BenShoshan M, Neuman S, Levhar N, Bubis M, et al. Individualized Dynamics in the Gut Microbiota Precede Crohn's Disease Flares. The American journal of gastroenterology. 2019;114(7):1142-51.
[526] Takenaka K, Ohtsuka K, Fujii T, Negi M, Suzuki K, Shimizu H, et al. Development and Validation of a Deep Neural Network for Accurate Evaluation of Endoscopic Images From Patients With Ulcerative Colitis. Gastroenterology. 2020;158(8):2150-7.
[527] Morell Miranda P, Bertolini F, Kadarmideen H. Investigation of gut microbiome association with inflammatory bowel disease and depression: a machine learning approach [version 2; peer review: 2 approved with reservations]. F1000Research. 2019;7(702).
[528] Waljee AK, Lipson R, Wiitala WL, Zhang Y, Liu B, Zhu J, et al. Predicting Hospitalization and Outpatient Corticosteroid Use in Inflammatory Bowel Disease Patients Using Machine Learning. Inflamm Bowel Dis. 2017;24(1):45-53.
[529] Jain S, Kedia S, Sethi T, Bopanna S, Yadav DP, Goyal S, et al. Predictors of long-term outcomes in patients with acute severe colitis: A northern Indian cohort study. J Gastroenterol Hepatol. 2018;33(3):615-22.
[530] Firouzi F, Rashidi M, Hashemi S, Kangavari M, Bahari A, Daryani NE, et al. A decision tree-based approach for determining low bone mineral density in inflammatory bowel disease using WEKA software. European journal of gastroenterology & hepatology. 2007;19(12):1075-81.

[531] Dorofeyev AE, Holub SV, Babayeva GH, Ananiin O. APPLICATION OF INTELLECTUAL MONITORING INFORMATION TECHNOLOGY IN DETERMINING THE SEVERITY OF THE CONDITION OF PATIENTS WITH INFLAMMATORY BOWEL DISEASES. Wiad Lek. 2021;74(3 cz 1):481-6.
[532] Gutierrez Becker B, Arcadu F, Thalhammer A, Gamez Serna C, Feehan O, Drawnel F, et al. Training and deploying a deep learning model for endoscopic severity grading in ulcerative colitis using multicenter clinical trial data. Ther Adv Gastrointest Endosc. 2021;14:2631774521990623.
[533] Li X, Liang D, Meng J, Zhou J, Chen Z, Huang S, et al. Development and Validation of a Novel Computed-Tomography Enterography Radiomic Approach for Characterization of Intestinal Fibrosis in Crohn's Disease. Gastroenterology. 2021;160(7):2303-16.e11.
[534] Yao H, Najarian K, Gryak J, Bishu S, Rice MD, Waljee AK, et al. Fully automated endoscopic disease activity assessment in ulcerative colitis. Gastrointest Endosc. 2021;93(3):728-36.e1.
[535] Gottlieb K, Requa J, Karnes W, Chandra Gudivada R, Shen J, Rael E, et al. Central Reading of Ulcerative Colitis Clinical Trial Videos Using Neural Networks. Gastroenterology. 2021;160(3):710-9.e2.
[536] Wang J, Ortiz C, Fontenot L, Xie Y, Ho W, Mattai SA, et al. High circulating elafin levels are associated with Crohn's disease-associated intestinal strictures. PLoS One. 2020;15(4):e0231796.
[537] Popa IV, Burlacu A, Mihai C, Prelipcean CC. A Machine Learning Model Accurately Predicts Ulcerative Colitis Activity at One Year in Patients Treated with Anti-Tumour Necrosis Factor α Agents. Medicina (Kaunas). 2020;56(11).
[538] Biasci D, Lee JC, Noor NM, Pombal DR, Hou M, Lewis N, et al. A blood-based prognostic biomarker in IBD. Gut. 2019;68(8):1386-95.
[539] Mohapatra S, Nayak J, Mishra M, Pati GK, Naik B, Swarnkar T. Wavelet Transform and Deep Convolutional Neural Network-Based Smart Healthcare System for Gastrointestinal Disease Detection. Interdiscip Sci. 2021;13(2):212-28.
[540] Takenaka K, Ohtsuka K, Fujii T, Oshima S, Okamoto R, Watanabe M. Deep Neural Network Accurately Predicts Prognosis of Ulcerative Colitis Using Endoscopic Images. Gastroenterology. 2021;160(6):2175-7.e3.
[541] Bossuyt P, Nakase H, Vermeire S, de Hertogh G, Eelbode T, Ferrante M, et al. Automatic, computer-aided determination of endoscopic and histological inflammation in patients with mild to moderate ulcerative colitis based on red density. Gut. 2020;69(10):1778-86.
[542] Waljee AK, Joyce JC, Wang S, Saxena A, Hart M, Zhu J, et al. Algorithms outperform metabolite tests in predicting response of patients with inflammatory bowel disease to thiopurines. Clinical gastroenterology and hepatology : the official clinical practice journal of the American Gastroenterological Association. 2010;8(2):143-50.
[543] Han L, Maciejewski M, Brockel C, Gordon W, Snapper SB, Korzenik JR, et al. A probabilistic pathway score (PROPS) for classification with applications to inflammatory bowel disease. Bioinformatics. 2018;34(6):985-93.
[544] Wang L, Fan R, Zhang C, Hong L, Zhang T, Chen Y, et al. Applying Machine Learning Models to Predict Medication Nonadherence in Crohn's Disease Maintenance Therapy. Patient preference and adherence. 2020;14:917-26.
[545] Pal LR, Kundu K, Yin Y, Moult J. CAGI4 Crohn's exome challenge: Marker SNP versus exome variant models for assigning risk of Crohn disease. Hum Mutat. 2017;38(9):1225-34.
[546] Doherty MK, Ding T, Koumpouras C, Telesco SE, Monast C, Das A, et al. Fecal Microbiota Signatures Are Associated with Response to Ustekinumab Therapy among Crohn's Disease Patients. mBio. 2018;9(2).
[547] Daneshjou R, Wang Y, Bromberg Y, Bovo S, Martelli PL, Babbi G, et al. Working toward precision medicine: Predicting phenotypes from exomes in the Critical Assessment of Genome Interpretation (CAGI) challenges. Hum Mutat. 2017;38(9):1182-92.
[548] Waljee AK, Liu B, Sauder K, Zhu J, Govani SM, Stidham RW, et al. Predicting corticosteroid-free endoscopic remission with vedolizumab in ulcerative colitis. Aliment Pharmacol Ther. 2018;47(6):763-72.

[549] Tong Y, Lu K, Yang Y, Li J, Lin Y, Wu D, et al. Can natural language processing help differentiate inflammatory intestinal diseases in China? Models applying random forest and convolutional neural network approaches. BMC Med Inform Decis Mak. 2020;20(1):248.
[550] McDonnell M, Harris RJ, Borca F, Mills T, Downey L, Dharmasiri S, et al. High incidence of glucocorticoid-induced hyperglycaemia in inflammatory bowel disease: metabolic and clinical predictors identified by machine learning. BMJ Open Gastroenterol. 2020;7(1).
[551] Jiang P, Wu S, Luo Q, Zhao XM, Chen WH. Metagenomic Analysis of Common Intestinal Diseases Reveals Relationships among Microbial Signatures and Powers Multidisease Diagnostic Models. mSystems. 2021;6(3).
[552] Waljee AK, Sauder K, Patel A, Segar S, Liu B, Zhang Y, et al. Machine Learning Algorithms for Objective Remission and Clinical Outcomes with Thiopurines. Journal of Crohn's & colitis. 2017;11(7):801-10.
[553] Isakov O, Dotan I, Ben-Shachar S. Machine Learning-Based Gene Prioritization Identifies Novel Candidate Risk Genes for Inflammatory Bowel Disease. Inflamm Bowel Dis. 2017;23(9):1516-23.
[554] Yu S, Chakrabortty A, Liao KP, Cai T, Ananthakrishnan AN, Gainer VS, et al. Surrogate-assisted feature extraction for high-throughput phenotyping. J Am Med Inform Assoc. 2017;24(e1):e143-e9.
[555] Ozcift A, Gulten A. Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. Computer Methods and Programs in Biomedicine. 2011;104(3):443-51.
[556] Cohen S, Dagan N, Cohen-Inger N, Ofer D, Rokach L. ICU Survival Prediction Incorporating Test-Time Augmentation to Improve the Accuracy of Ensemble-Based Models. IEEE Access. 2021;9:91584-92.
[557] Feng B-J. PERCH: A Unified Framework for Disease Gene Prioritization. Human Mutation. 2017;38(3):243-51.
[558] Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. The American Journal of Human Genetics. 2016;99(4):877-85.
[559] Tian Y, Pesaran T, Chamberlin A, Fenwick RB, Li S, Gau C-L, et al. REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. Scientific Reports. 2019;9(1):12752.
[560] Broad Institute of MIT. DRAGEN-GATK Update: Let's get more specific 2022 [Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360039984151-DRAGEN-GATK-Update-Let-s-get-more-specific.
[561] Broad Institute of MIT. DRAGEN-GATK 2022 [Available from: https://gatk.broadinstitute.org/hc/en-us/articles/360045944831-DRAGEN-GATK.
[562] O'Driscoll A, Daugelaite J, Sleator RD. 'Big data', Hadoop and cloud computing in genomics. Journal of Biomedical Informatics. 2013;46(5):774-81.
[563] UK B. Final data release from the world's largest whole exome sequencing project 2022 [Available from: https://www.ukbiobank.ac.uk/learn-more-about-uk-biobank/news/final-data-release-from-the-world-s-largest-whole-exome-sequencing-project.
[564] Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q. Opportunities and challenges in long-read sequencing data analysis. Genome Biology. 2020;21(1):30.
[565] Hallock H, Marshall SE, t Hoen PAC, Nygård JF, Hoorne B, Fox C, et al. Federated Networks for Distributed Analysis of Health Data. Front Public Health. 2021;9:712569-.
[566] Lancaster O, Beck T, Atlan D, Swertz M, Thangavelu D, Veal C, et al. Cafe Variome: general-purpose software for making genotype-phenotype data discoverable in restricted or open access contexts. Hum Mutat. 2015;36(10):957-64.
[567] Rosen MJ, Dhawan A, Saeed SA. Inflammatory Bowel Disease in Children and Adolescents. JAMA Pediatrics. 2015;169(11):1053-60.
[568] Magg T, Shcherbina A, Arslan D, Desai MM, Wall S, Mitsialis V, et al. CARMIL2 Deficiency Presenting as Very Early Onset Inflammatory Bowel Disease. Inflammatory Bowel Diseases. 2019.

[569] Lehle AS, Farin HF, Marquardt B, Michels BE, Magg T, Li Y, et al. Intestinal Inflammation and Dysregulated Immunity in Patients With Inherited Caspase-8 Deficiency. Gastroenterology. 2019;156(1):275-8.
[570] Zhou SM, Fernandez-Gutierrez F, Kennedy J, Cooksey R, Atkinson M, Denaxas S, et al. Defining disease phenotypes in primary care electronic health records by a machine learning approach: A case study in identifying rheumatoid arthritis. PLoS ONE. 2016;11 (5) (e0154515).
[571] Schwaighofe A, Tresp V, Mayer P, Krause A, Beuthan J, Rost H, et al. Classification of rheumatoid joint inflammation based on laser imaging. IEEE Transactions on Biomedical Engineering. 2003;50(3):375-82.
[572] Douglas GM, Hansen R, Jones CM, Dunn KA, Comeau AM, Bielawski JP, et al. Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. Microbiome. 2018;6 (1) (13).
[573] Jain S, Kedia S, Sethi T, Bopanna S, Yadav D, Goyal S, et al. Predictors of long-term outcomes in patients with acute severe ulcerative colitis: A northern Indian cohort study. Gastroenterology. 2017;152 (5 Supplement 1):S372.
[574] Reddy BK, Delen D, Agrawal RK. Predicting and explaining inflammation in Crohn's disease patients using predictive analytics methods and electronic medical record data. Health Informatics J. 2019;25(4):1201-18.
[575] Maeda Y, Kudo SE, Mori Y, Misawa M, Ogata N, Sasanuma S, et al. Fully automated diagnostic system with artificial intelligence using endocytoscopy to identify the presence of histologic inflammation associated with ulcerative colitis (with video). Gastrointest Endosc. 2019;89(2):408-15.

Bibliography

Papers published during this PhD candidature (since September 2018)

1. **Stafford IS**, Mossotto E, Ashton JJ, Cheng G, Beattie RM, Ennis S. Supervised machine learning classifies inflammatory bowel disease patients by subtype using whole exome sequencing data. *J Crohns Colitis*. May 2023. https://doi.org/10.1093/ecco-jcc/jjad084
2. Ashton JJ, Cheng G, **Stafford IS**, Kellermann M, Seaby E, Cummings FJR, Coelho TF, Batra A, Afzel NA, Beattie RM, Ennis S. Prediction of Crohn's disease stricturing phenotype using a NOD2-derived genomic biomarker. *Inflamm. Bowel Dis*. Apr 2023. DOI: 10.1093/ibd/izac205
3. Mossotto E, Boberska J, Ashton JJ, **Stafford IS**, Cheng G, Baker J, Borca F, Phan HTT, Coelho TF, Beattie RM, Claus SP, Ennis S. Evidence of a genetically driven metabolomic signature in actively inflamed Crohn's disease. *Sci Rep*. Aug 2022. https://doi.org/10.1038/s41598-022-18178-9
4. **Stafford IS**, Gosink MM, Mossotto E, Ennis S, Hauben M. A systematic review of artificial intelligence and machine learning applications to inflammatory bowel disease, with practical guidelines for interpretation. *Inflamm. Bowel Dis*. June 2022. https://doi.org/10.1093/ibd/izac115
5. Ashton JJ, Boukas K, **Stafford IS**, Cheng G, Haggarty R, Coelho TAF, Batra A, Afzal NA, Williams AP, Polak ME, Beattie RM, Ennis S. Deleterious Genetic Variation Across the NOD Signaling Pathway Is Associated With Reduced NFKB Signaling Transcription and Upregulation of Alternative Inflammatory Transcripts in Pediatric Inflammatory Bowel Disease. *Inflamm Bowel Dis*. Jun 2022. doi:10.1093/ibd/izab318
6. Coelho TF, Sonnenberg-Riethmacher E, Gao Y, Mossotto E, Khojanazarov A, Griffin A, Mukanova S, Ashimkhanova A, Haggarty R, Borissenko A, Ashton JJ, **Stafford IS**, Batra A, Afzal NA, Stanton MP, Vadgama B, Adrisova K, Beattie RM, Williams AP, Ennis S, Riethmacher D. Expression profile of the matricellular protein periostin in paediatric inflammatory bowel disease. *Sci Rep*. Mar 2021 Mar. doi: 10.1038/s41598-021-85096-7.
7. Ashton JJ, Boukas K, Davies J, **Stafford IS**, Vallejo AF, Haggarty R, Coelho TAF, Batra A, Afzal NA, Vadgama B, Williams AP, Beattie RM, Polak ME, Ennis S. Ileal Transcriptomic Analysis in Paediatric Crohn's Disease Reveals IL17- and NOD-signalling Expression Signatures in Treatment-naïve Patients and Identifies Epithelial Cells Driving Differentially Expressed Genes, *J Crohns Colitis*. May 2021. doi: 10.1093/ecco-jcc/jjaa236.

8. Coelho TF, Mossotto E, Gao Y, Haggarty R, Ashton JJ, Batra A, **Stafford IS**, Beattie RM, Williams AP, Ennis S. Immunological Profiling of Paediatric Inflammatory Bowel Disease Using Unsupervised Machine Learning. *J Pediatr Gastroenterol Nutr*. Jun 2020. doi: 10.1097/MPG.0000000000002719
9. **Stafford IS**, Kellermann M, Mossotto E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ Digit Med*. Mar 2020. doi: 10.1038/s41746-020-0229-3.
10. Ashton JJ, Mossotto E, **Stafford IS**, Haggarty R, Coelho TAF, Batra A, Afzal NA, Mort M, Bunyan D, Beattie RM, Ennis S. Genetic Sequencing of Pediatric Patients Identifies Mutations in Monogenic Inflammatory Bowel Disease Genes that Translate to Distinct Clinical Phenotypes. *Clin Transl Gastroenterol*. Feb 2020. doi: 10.14309/ctg.0000000000000129