Big data in IBD: big progress for clinical practice

Nasim Sadat Seyed Tabib ,[1] Matthew Madgwick,[2,3] Padhmanand Sudhakar,[1,2,3] Bram Verstockt ,[4,5] Tamas Korcsmaros,[2,3] Séverine Vermeire[1,5]

[1]Department of Chronic Diseases, Metabolism and Ageing, TARGID, KU Leuven, Leuven, Belgium
[2]Organisms and Ecosystems, Earlham Institute, Norwich, UK
[3]Gut microbes in health and disease, Quadram Institute Bioscience, Norwich, UK
[4]Translational Research in GastroIntestinal Disorders, KU Leuven, Leuven, Belgium
[5]Department of Gastroenterology and Hepatology, KU Leuven University Hospitals Leuven, Leuven, Belgium

Correspondence to
Dr Séverine Vermeire, Department of Chronic Diseases, Metabolism and Ageing - TARGID, KU Leuven, Leuven B-3000, Belgium; severine.vermeire@uzleuven.be

Received 11 October 2019
Revised 5 February 2020
Accepted 6 February 2020
Published Online First 28 February 2020

To cite: Seyed Tabib NS, Madgwick M, Sudhakar P, *et al*. *Gut* 2020;**69**:1520–1532.

ABSTRACT

IBD is a complex multifactorial inflammatory disease of the gut driven by extrinsic and intrinsic factors, including host genetics, the immune system, environmental factors and the gut microbiome. Technological advancements such as next-generation sequencing, high-throughput omics data generation and molecular networks have catalysed IBD research. The advent of artificial intelligence, in particular, machine learning, and systems biology has opened the avenue for the efficient integration and interpretation of big datasets for discovering clinically translatable knowledge. In this narrative review, we discuss how big data integration and machine learning have been applied to translational IBD research. Approaches such as machine learning may enable patient stratification, prediction of disease progression and therapy responses for fine-tuning treatment options with positive impacts on cost, health and safety. We also outline the challenges and opportunities presented by machine learning and big data in clinical IBD research.

INTRODUCTION

Precision medicine holds great promise to improve the landscape of IBD course of care for an individual patient, providing the most beneficial therapy while minimising the risk. The ultimate goals of precision medicine include stratifying patients based on disease subtypes and severity, disease progression and treatment response using personal and clinical data coupled with molecular profiling data of patients.[1 2] IBD, with its two main subtypes, Crohn's disease (CD) and UC, is a complex inflammatory disease with a wide range of contributing factors including host genetics, immune system, environmental exposures and the gut microbiome.[3–5] The inherent complexity of the disease introduces a large number of confounding factors, which stand in the way of accurate diagnosis and precision medicine.[6]

The term 'big data' is generally referred to as large volume of rapidly produced data from variable sources, known as the three 'V's (volume, velocity and variety).[7] Over the past decades, the production and availability of data that could inform healthcare has increased remarkably mainly due to technological advancements and falling costs of data generation. Most important sources of data in IBD comprise study cohorts, clinical trials, administrative and electronic health record databases, patient-reported outcomes databases, medical imaging databases and omics datasets (including genomics, transcriptomics, proteomics and metabolomics, as well as environmental omics) (figure 1). The use of big data in IBD allows medical researchers to reveal disease-related trends, associations and patterns to propel our understanding of IBD forward and to inform clinical practice.[2 8] However, due to the high complexity of big data and the long list of confounding factors, interpreting these data is not trivial and warrants approaches that can uncover hidden patterns in these large and complex datasets.[9]

Recent developments in computational biology have driven the integration of big data and molecular networks using the principles of systems biology and machine learning. Systems biology centres around the holistic and mathematical modelling of complex biological system.[10] Machine learning is a subset of artificial intelligence, which refers to the ability of algorithms to learn from data in order to detect patterns and make decisions (without explicitly being programmed what to do) (Box 1).[11] Machine learning algorithms provide the means and opportunity to investigate large amounts of data and thus help identify patterns behind complex medical conditions. These analytical approaches allow categorisation of patients based on their specific differences through screening a patient's genome, transcriptome, proteome, epigenome, immunome and microbiome. Integrating the omics datasets using systems biology-based approaches may advance understanding of the underlying causative factors in individual patients. The arrival of systems biology and machine learning into IBD clinical research has allowed researchers to capture complex associations andincreased understanding of disease mechanisms in IBD. In this narrative review, we provide an overview of the sources of big data in IBD. We discuss how artificial intelligence could help us better understand IBD pathogenesis and how some components of it have already begun to shape our knowledge of IBD. We address how artificial intelligence could contribute to the diagnosis and prognosis of IBD, and whether it could assist with predictions of therapy efficacy and adverse effects. As a final point, we argue the potential that artificial intelligence provides for personalised medicine in IBD and evaluate the feasibility of big data in IBD disease management.

ROLE OF MACHINE LEARNING AND SYSTEMS BIOLOGY IN THE INTERPRETATION OF BIG DATA IN IBD RESEARCH

The main challenge faced by many scientists is to extract meaningful information through integrating different sources of data and thereby discover disease association patterns. Classical statistical methods are not powerful enough to explain the

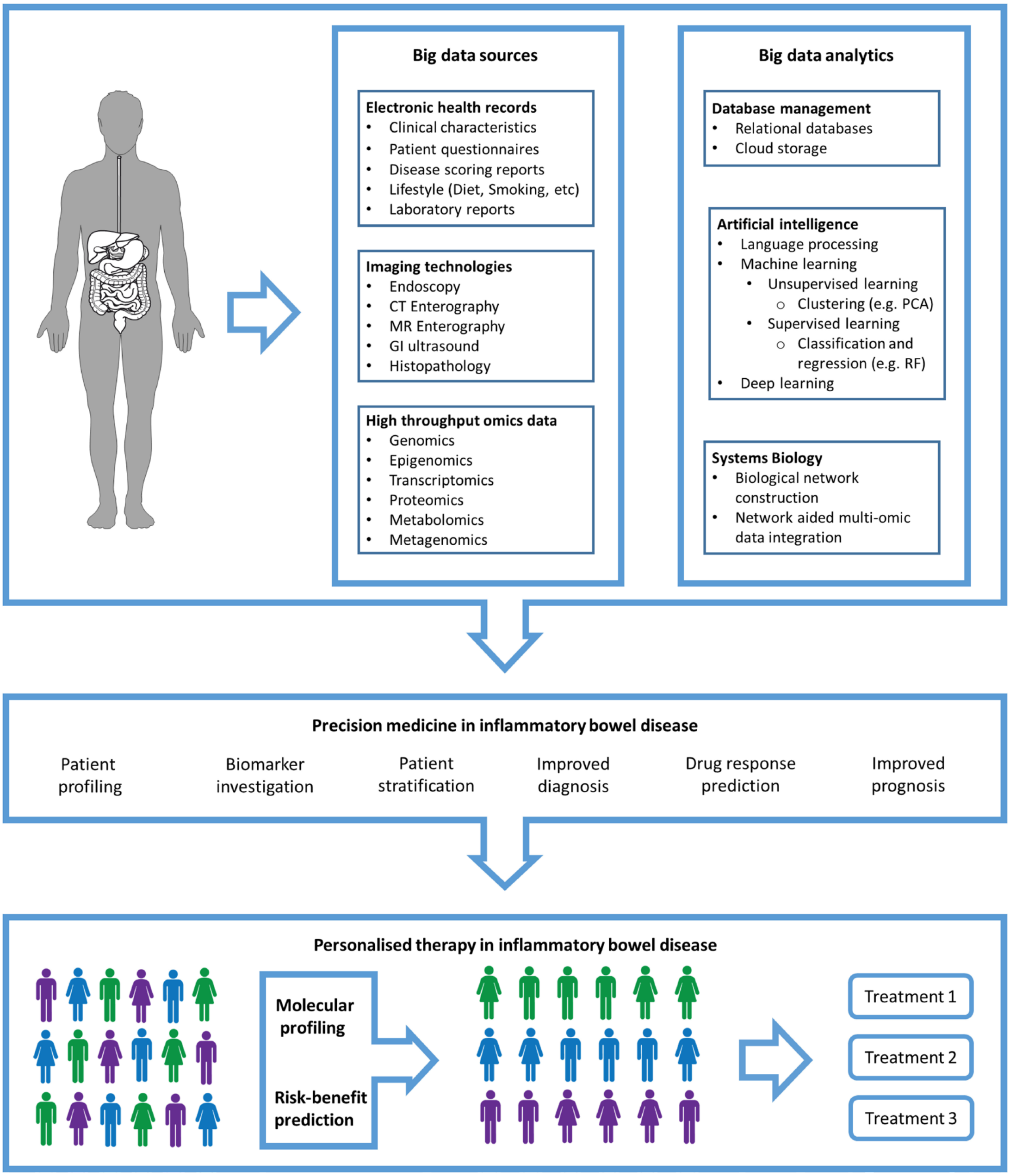

Figure 1 Precision medicine in IBD. Generation of big data from thousands of individuals, along with analytical advancements such as machine learning and systems biology, assists the application of precision medicine and therefore allows patient stratification for personalised therapeutic intervention and disease management strategies. MR, magnetic resonance; PCA, principal component analysis; RF, random forest.

underlying milieu of pathogenic and causative factors in IBD. Hence, scientists have adopted different analytical methodologies. Generally, such analytical methodologies are categorised into two main groups, namely, systems biology and machine learning. These are more powerful and flexible methods in biomedical data science and have the potential to uncover novel insights into disease pathogenesis.[12 13]

Systems biology paves the way for data integration and analysis from a functional perspective, and it has assisted in identifying the pathophysiological mechanisms of IBD. The approach of systems biology typically involves the use of networks (mostly molecular networks such as protein–protein interaction networks, regulatory networks involving transcription factors and metabolic networks) to capture the physical and signalling interactions and to interpret contextual measurements such as expression of genes, proteins and metabolites. This approach thereby provides a framework to identify key components and/or pathways which mediate the pathogenesis of the disease. Brooks *et al* identified different clusters of patients with UC using network footprints created by combining mutation data, protein–protein interaction networks and gene expression data.[14]

In the past decade, machine learning has attracted much attention from groups engaged in IBD research, owing to its ability to learn complex patterns and make prediction. With machine learning as a framework, several attempts have been made to use different types of omics and clinical datasets to improve our understanding of disease mechanism. Given that omics datasets, such as RNAseq data, comprise expression information of thousands of genes (features) with far few samples, feature selection is of great importance. Machine learning algorithms take advantage of data-dependent automatic feature learning, while systems biology approaches need to be manually programmed. Machine learning algorithms can learn how to integrate several predictors to identify a representative subset of input data[15]; for example, a machine learning algorithm using the concept of random forests identified a panel of 50 faecal bacteria capable of distinguishing active and remission states in patients with CD.[16]

Box 1

Artificial intelligence terminology

Artificial intelligence: the field of computer science which concerns the theory and development of computers to perform tasks which usually requires human intelligence, such as imagine classification, speech recognition and decision-making.
Machine learning: a field of artificial intelligence which refers to the computers' ability to learn to make decisions or detect patterns (without explicitly being programmed) from data.
Deep learning: a subfield of machine learning that exploits many layers of non-linear information processing for supervised or unsupervised feature extraction and transformation, and for pattern analysis and classification using various neural network frameworks.
Supervised learning: the task of an algorithm learning a function that maps an input to an output based on provided example data.
Unsupervised learning: the task of a machine learning algorithm to learn the underlying data structure of unlabeled example data, for example, finding commonalities, leading to insights and therefore a greater understanding of the example data.
Classification: the process of predicting a class/subcategory of given data points from known example data.
Generalisation: refers to how well the machine learning model learns the underlying data and the model's ability to apply this to specific examples not seen by the model during training.
Ensemble learning: the union of homologous or heterogeneous machine learning algorithms whose predictions are combined to achieve greater performance than just the individual machine learning algorithm could achieve alone.
Support vector machine: this is a discriminative classifier which determines classes from a separating hyperplane. Through the use of a kernel, SVMs can be adapted to suit non-linear problems.
Random forests: a homologous ensemble algorithm which constructs a great number of decision trees at training.
Matrix factorisation: an algorithm which extracts meaningful association from an incomplete data matrix and transforms them in a lower dimensional latent space, also known as recommender systems.

Genomics

IBD is considered as a polygenic disease, with the exception of rare monogenic cases.[17] The notable example of research into the genetic basis of IBD is the introduction of *NOD2* as the first CD susceptibility gene.[18] To date, the continued search for genetic determinants of IBD identified 242 variants associated with IBD,[17] of which 45 have been fine mapped to statistically significant causal variants. Interestingly, associated regions indicate that there is a profound overlap between IBD and other immune-mediated inflammatory diseases. However, merely a small percentage of heritability is explained by the identified loci.[17]

To further resolve the genetic architecture of IBD, machine learning and data integration could be employed to propel the gene discoveries. The main issue with association studies is the imbalance between the number of patient samples and the number of single-nucleotide polymorphisms (SNPs) that are being analysed. In addition, the classical genotype–phenotype association at high statistical confidence neglects a considerable fraction of genetic variation. Machine learning could be used to detect meaningful patterns containing thousands of DNA variants, regardless of the statistical significance level.[19–22] This could result in predictions of genetic markers and variants with greater accuracy. An exemplary study was conducted using data from the International IBD Genetics Consortium's Immunochip project. To reduce the number of SNPs, Wei and colleagues applied a less rigid statistical confidence limit (p values of $<10^4$ and minor allele frequency of <0.01) followed by a machine learning classifier-based feature selection method (the penalised logistic regression model). The authors defined 573 SNP-based CD and 366 SNP-based UC predictive models with superior area under the receiver operating characteristic curve (AUC) values than the log OR-based models (AUCs of 0.86 (95% CI 0.85 to 0.86) and 0.826 (95% CI 0.81 to 0.83) for CD and UC, respectively).[23] Another interesting study was conducted using the UK Inflammatory Bowel Disease Genetics Consortium and UK10K consortium for the controls, which cumulatively comprises approximately 8000 individuals (4280 patients and 3652 controls). In this study, a machine learning model, a support vector machine (SVM), was used to hunt for novel genetic variants, which resulted in the identification of a missense variant in *ADCY7* associated with UC with a frequency of 0.6%.[24] A recent study reanalysed the Immunochip dataset using different machine learning models, including random forests and neural networks. Romagnoni *et al* identified new variants with minor effects, in addition to almost all of the previously known variants among the best predictors of CD.[25]

Advancements in sequencing technologies allow a more in-depth genomic screening. Scientists have used whole genome/exome sequencing particularly to discover rare genetic variants, such as *NOX1*, contributing to very early-onset IBD.[26] Machine learning methodologies, particularly deep learning, are resourceful tools for not only making predictions but also extracting biomedical insights.[27–30] In a notable publication, Zou *et al* provided a primer on deep learning for genomic data analysis accompanied with practical guidelines for the discovery of DNA-binding motifs.[31]

Transcriptomics and proteomics

Investigating the downstream effects of genomic aberrations, namely, on the transcriptome and proteome, provides additional molecular details to unravelling IBD pathogenesis. Differential gene expression analysis has been used to identify key genes and pathways underlying IBD pathogenesis. Transcriptomic analyses of human ileum and colonic samples have helped to uncover the roles of different pathways driving inflammation in IBD. For example, inflamed and non-inflamed tissues have altered gene expression in CD and UC. To investigate the functional significance of these modifications and to characterise their molecular signatures in colonic tissue, an integrated systems approach has highlighted significant enrichment in proteasome and apoptosis pathways.[32] With protein–protein interaction network analysis, Li *et al* identified *MAPK3*, *NDRG1* and *HLA-DRA* as key players in disease pathogenesis. Following a similar approach, Hong *et al* identified altered gene expression profiles and key cellular pathways in patients with inflamed and non-inflamed intestinal mucosa with CD, including immune response, chemokine signalling and cell adhesion.[33]

Weighted gene coexpression network analysis allows researchers to detect genes that are upregulated or downregulated in tandem under specific conditions.[34 35] For example, Lin *et al* revealed important pathogenic roles for *IL-8* and *MMP-9* in the colonic tissues of patients with UC by combining gene coexpression and protein–protein interaction networks.[36] A similar study in the context of gene expression alteration in different stages of CD by Verstockt *et al* pinpointed that dysregulation of the coexpression network is more evident in newly diagnosed and late-stage CD compared with recurrent CD.[37] Likewise, this network approach can elucidate biological mechanisms driving treatment resistance to biological therapies, such as with tumour necrosis factor (TNF) inhibitor agents.[38] Another functional approach to explore the gene expression data is metabolism-level interpretation using Recon 2,[39] the model of the human metabolic network. Using this model, critical pathways such as cellular transport of thiamine and bile acid metabolism have been identified.[40]

Yuan and colleagues reported 41 discriminatory IBD-related genes by combining machine learning and systems biology. In searching for novel candidate genes, the authors used a two-step feature selection on microarray data from patients with CD, UC and control individuals. First, they ranked thousands of genes according to their correlation to diagnosis and the redundancies between each gene related to all other genes in the ranked list. Then, using an SVM as a machine learning classifier, they identified a feature set containing 21 genes, which yield the highest prediction accuracy. Additionally, based on the concept of functional similarity among closely related proteins, the authors used the protein–protein interaction network of the proteins encoded by those 21 genes and applied the shortest path approach (typically defined as the path with the least number of links between two proteins in a network) to find an additional 20 candidate genes.[41] In another interesting study by Isakov *et al*, novel candidate genes were identified by developing a machine learning model trained on expression values of known IBD susceptibility genes and their functional annotations. The authors used the feature importance of a machine learning classifier as the feature selection method.[42]

Environmental 'omics'

The gut microbiota, which comprise intestinal bacteria, fungi, archaea and viruses, is an essential part of the human GI tract and plays a pivotal role in human health. In homeostatic condition, there is a state of immunological tolerance to the commensal intestinal microbiota. It has been established that perturbation of composition, function and structure of the gut microbiota, known as dysbiosis, is one of the key players in IBD pathogenesis.[43] However, it is still not clear whether this dysbiosis is the cause or consequence in patients with IBD.

There is a decline in both species diversity and richness in patients with IBD. Several studies have reported an increase in the abundance of certain species from the Proteobacteria phylum, such as *Escherichia coli*, and a decline in anti-inflammatory butyrate-producing bacteria species, such as *Faecalibacterium prausnitzii*, belonging to the Firmicutes phylum. Additionally, a longitudinal study suggested an increase in dynamic fluctuation of the gut microbiome composition in patients with IBD.[44]

Much less is currently known on the role played by viruses in the dysbiotic state in patients with IBD. Recent advances in sequencing technologies and data analytic techniques have enabled in-depth characterisation of microbiota communities to investigate IBD pathogenesis using meta-level omics datasets, namely, metagenomics, metatranscriptomics, metaproteomics and metabolomics. Deep metagenomics paved the way to study gut resident fungi, archaea and viruses in both healthy and disease states. Different stool virome profiles have been observed in patients with IBD compared with healthy individuals.[45] Zuo and colleagues used machine learning-based clustering to define viral metacommunities in rectal mucosa derived from patients with UC. The predominant viral community among patients with UC showed decreased viral diversity, richness and evenness, particularly among *Caudovirales* species. However, two species of *Caudovirales* (*Escherichia phage* and *Enterobacteria phage*) were much more common among patients with UC compared with healthy controls. This suggests a loss of corelationship between the viruses and bacteria, which can cause microbiota dysbiosis and intestinal inflammation.[46]

The interplay between the microbial composition and metabolism of the gut is an interesting nexus in IBD. While much of the previous research on this interaction level has been interpretive in nature, most of the studies on the gut protein and metabolic composition used shotgun metagenomic technique. Thus, by comparing the abundance of enzymatic genes across samples, scientists have been able to infer the effect of variations in microbial composition on the protein and metabolic levels. An example of this is the study by Greenblum *et al* in which they used faecal metagenomics to build metabolic networks. They demonstrate topological differences by which IBD-associated metabolic networks interact with the gut environment and the host.[47] There is a growing number of investigations applying the approaches of metaproteomics and metabolomics. Particularly, there are two avenues in which metaproteomics-based investigations have been employed, the mucosal–luminal interface analysis and the stool metaproteome profiling. Li *et al* investigated the protein co-occurrence network at the mucosal surface of six different colonic regions. Employing weighted correlation network analysis and multiple clustering methods such as hierarchical clustering, they identified distinct functional protein modules (protein clusters that alter together) in association with non-IBD, UC and CD disease states.[48] In addition to systems biology methods, machine learning could be applied to define relevant protein clusters. Profiling of stool samples revealed that metaproteomic signatures in patients with CD differ from those of healthy individuals. By integrating metagenomics and metaproteomics, and applying a hierarchical clustering method, Erickson *et al* reported a depletion of several microbial proteins in patients with CD with ileal involvement, such as proteins in the butyrate pathway which corresponded to a reduction in the Firmicutes phylum.[49]

Multiomics data integration

In more recent investigations, researchers have been collecting different levels of omics data from patients with IBD to investigate the crosstalk between the key players in IBD pathogenesis. An interesting area in which multiomics data integration has been applied is to characterise the dysregulated multifaceted interactions between various host and microbial factors in IBD. For example, Häsler *et al* studied the transition of intestinal homeostasis to dysbiosis by integrating multiple levels of data, namely, the mucosal transcriptomic, post-transcriptional alterations and the mucosal microbiome of patients with UC and CD in comparison with healthy individuals. The authors identified the enrichment of host transcript splicing events as a result of the interplay between microbial and host factors which probably mediate the transition of intestinal homeostasis to dysbiosis in patients with IBD.[50] In order to investigate the dysbiosis at the functional level, Lloyd-Price *et al* followed up 132 patients with IBD for 1 year and performed extensive molecular profiling of all patients. The authors revealed a distinctive upsurge in the ratio of facultative anaerobes to obligate anaerobes, along with disruptions at the molecular level, including microbial transcription division (within clostridia) and metabolite disruptions (acylcarnitines, bile acids, and short-chain fatty acids). Additionally, they reported noticeable alterations in the composition and function of microbiota with regard to different disease activity states.[51]

CURRENT PARADIGM OF IBD DISEASE MANAGEMENT AND ITS LIMITATIONS

The scope of IBD treatment is extending swiftly, with the introduction of new biologics and small molecules as a result of the improved understanding of the disease pathophysiology. With novel treatment options (targeting different aspects of IBD pathophysiology) such as anticytokine or chemokine agents, antiadhesion molecules, stem cell therapy and manipulation of the gut microbiota becoming increasingly available, it is time to move beyond the 'one-size-fits-all' approach.[52]

IBD management (figure 2) encompasses three different stages, starting with diagnosis, followed by the assessment of disease and the choice of therapy regimens, follow-up assessments and associated treatment changes, if necessary. Disease monitoring is key and is currently carried out by tracking different markers like faecal calprotectin, serum C reactive protein, also colonoscopy and/or medical imaging technologies such as abdominal ultrasound and MRI.[53 54] Hitherto, the clinical decision on the choice of therapeutic strategy depended on the response and tolerability of treatment in patients. However, in light of recent innovative therapies in IBD, a more accurate method is warranted to assist and complement existing management.[55] In recent years, there has been an increasing interest in the application of machine learning in IBD clinical research. Using machine learning for personalised predictions will not only strengthen medical care and improve outcomes but also considerably decrease healthcare expenditure. Despite the importance of health economics, there are little published data on the cost-effectiveness of artificial intelligence in healthcare. An interesting example is the study conducted by Bremer *et al*, who deployed a machine learning methodology to predict the individual outcome and

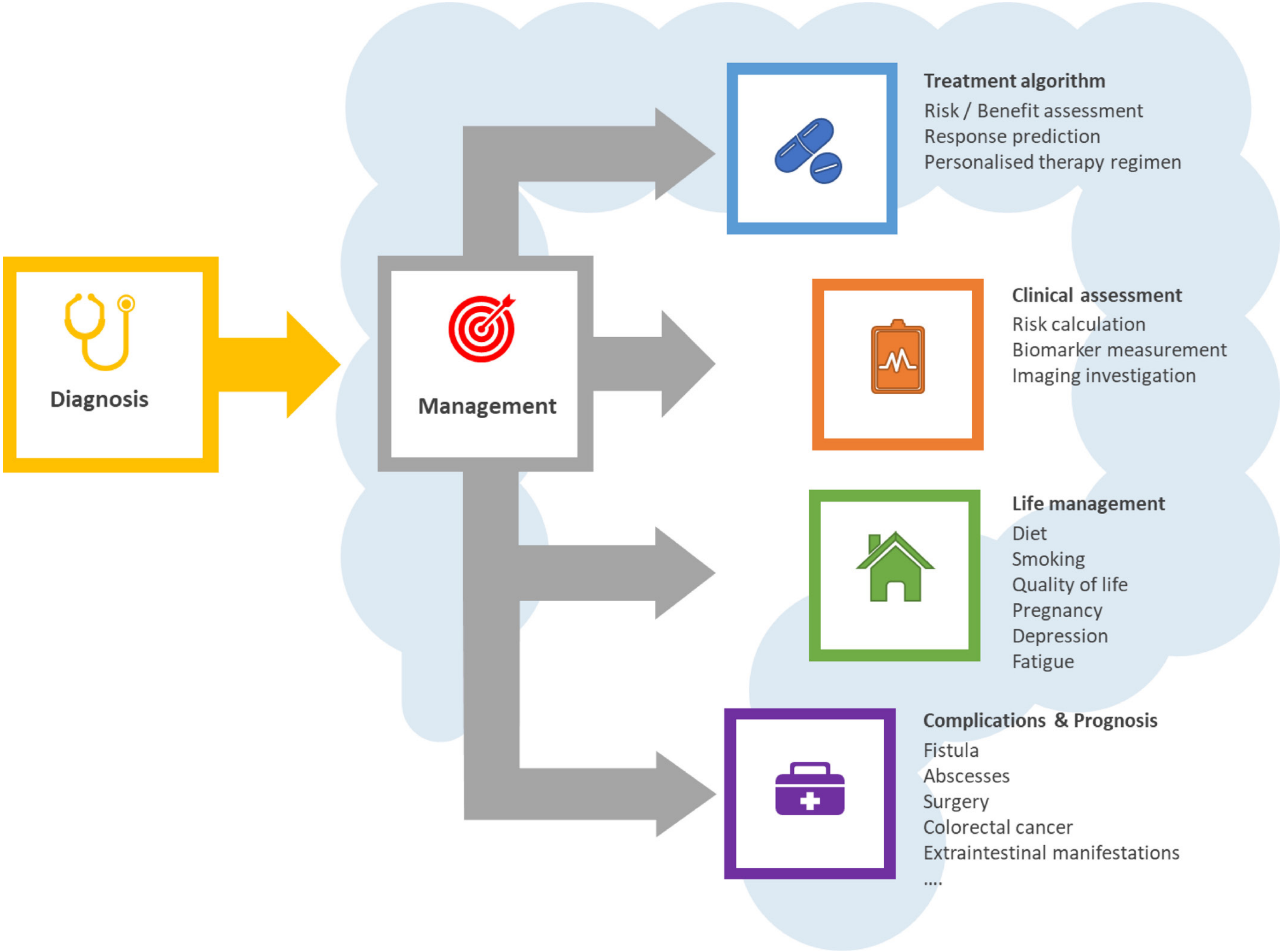

Figure 2 Clinical management of IBD from the point of diagnosis to life-term monitoring and follow-up. Each stage of the disease management process can potentially be subjected to precision medicine-aided improvement of patient care to reduce the socioeconomic burden on patients, clinicians and the healthcare system.

costs for patients with depressive disorders prior to the start of intervention in order to allocate patients to the most beneficial treatment.[56] In the field of gynaecology, Wang *et al* proposed a machine learning-based strategy for urinalysis which significantly increased the detection rate of the pathogen *Trichomonas vaginalis* in a cost-effective manner.[57]

Diagnosis and risk stratification

The current paradigm of IBD classification, which relies on invasive ileocolonoscopy and biopsies, does not adequately capture the broad spectrum of phenotypes of the disease or the patient-specific manifestations of its comorbid conditions. Recent research has focused on identifying and evaluating potential non-invasive diagnostic markers to diagnose IBD, differentiate it from other disorders and potentially improve its classification. There is great interest in the diagnostic value of genomics data, with over 240 IBD-associated risk loci already identified using genome wide association study (GWAS) data. A genotype–phenotype study associated three loci, *NOD2*, *MHC* and *MST*, with subphenotypes of IBD, particularly disease location.[58] Exome sequencing has arisen with the promise of unravelling the genetics of complex diseases. However, extracting disease-associated sequence variants is challenging due to inherited diversity of genomic variation. By incorporating exome sequence data with biological knowledge, such as functional interaction networks, into a matrix factorisation-based machine learning model, Jeong and Kim were able to distinguish patients with CD from healthy individuals (AUC=0.81).[59]

Likewise, molecular and cellular signatures can enable stratification of patients based on underlying pathways that drive their disease. Gene expression profiling is a major area of interest in the search of clinically associated signatures for IBD class prediction. To identify a set of genes distinguishing between UC and CD, novel machine learning-based methods have been used. Two examples which stand out are the PROPhet software package,[60] which automatically selects the best classifier and the optimal selection of genes to distinguish disease subtypes. Montero-Meléndez *et al* used this technique with microarray gene expression profiling of colonic biopsies to identify predictive transcriptional signatures associated with either CD or UC.[61] The second example is the Probabilistic Pathway Score, which is a pathway-based machine learning model that uses gene interactions to identify molecular pathways affected by the disease of interest and identify similarities and differences between them.[62] Proteomic signature is another promising nexus in biomarker research. Machine learning models have also been used with proteomic data to stratify patients with IBD. For example, Seeley *et al* investigated the protein signatures from colonic tissues using an SVM machine learning classifier trained on 25 peaks from histology-based mass spectrometry data. The model was able to discriminate patients with CD and UC from each other with an accuracy rate of 76.9%.[63] Another interesting area of biomarker research in IBD is microRNAs (miRNAs), a group of small noncoding RNA molecules which control gene expression and protein production and are detectable in many sources such as blood and urine. Hence, miRNAs hold great promise as potential non-invasive diagnostic markers. miRNAs are dysregulated in IBD.[64] Therefore, researchers have attempted to demonstrate the diagnostic value of circulating miRNAs signatures in the blood as diagnostic biomarkers using machine learning modelling, including random forests and SVM.[65 66]

An interesting example of exploring the diagnostic value of a set of biomarkers is the study conducted by Plevy *et al* combining genetic variants, serological and inflammatory markers to establish a diagnostic model to distinguish patients with IBD from those without IBD (healthy individuals or other diseases) and to separate patients with CD from UC. Based on the data from 1520 individuals, the authors selected 17 statistically significant markers and trained a random forest classifier, a machine learning algorithm, to differentiate the clinical groups.[67]

Machine learning approaches also hold great promise in unravelling disease-specific microbial signatures. Multiple machine learning-based microbiome frameworks have been established such as Multivariate Association with Linear Models (MaAsLin),[68] Metagenomic prediction Analysis based on Machine Learning[69] and phylogenetic convolutional neural networks[70] which incorporate patient clinical data, knowledge of microbial strains and knowledge of phylogenetic structure, respectively. Integrating additional information is expected to enhance the classification performance of microbiome-based machine learning models. As an example, Gevers *et al* were able to use rectal mucosa-associated microbiome signatures to distinguish paediatric patients with CD from patients with other GI tract conditions by integrating patient clinical data age, gender and past antibiotic use with the microbiome profiles using MaAsLin workflow.[71]

While the initial results of biomarker identification are promising, there is still a long way to go before these biomarkers can be applied in clinical practice, mainly due to the heterogeneity of the disease, diverse comorbidity factors and, importantly, lack of validation. The emergence of big data and big data analytics has led to a pile of studies and hypotheses. Although these approaches show great potential in a study-by-study basis, to translate these findings to a clinical setting, it is crucial to distinguish true discoveries from red herrings. Therefore, replication and validation studies in much larger cohort sizes are required. To achieve this, large and up-to-date clinical biobanks with a variety of different data types, including molecular, clinical and host characteristics, will be required to fully leverage these analytical methodologies. In precision medicine era, many national and international collaborative efforts are under way aimed at improving clinical research (figure 3).[72–84]

Advances in imaging technologies

Image recognition is one of the major applications of artificial intelligence, particularly deep learning, and holds great promise in assisting the fields of biological and medical imaging. Deep learning is a collection of algorithms in the field of machine learning with an outstanding ability to decode the contents of images. This has led to a proliferation of studies with an attempt to automate the interpretation and the evaluation of medical images, such as endoscopy, histopathology, and CT/MRI. Evaluating endoscopic inflammation, characterisation of lesions and assessment of mucosal healing is essential for proper management in IBD. However, endoscopic assessment of inflammation in IBD is highly subjective with high interobserver variability. Computer-aided scores would be much more objective for the interpretation of the endoscopic images.[85] For example, a deep learning-based model showed performance comparable to those of experienced gastroenterologists for the classification of endoscopic severity of UC into two groups: remission (Mayo 0 or 1 endoscopic score) and moderate to severe (Mayo 2 or 3 endoscopic score).[86] A novel objective computer-based score to assess UC disease activity based on endoscopic images has been developed. In particular, deep learning has been used to extract different layers of pixel data, such as measuring the redness

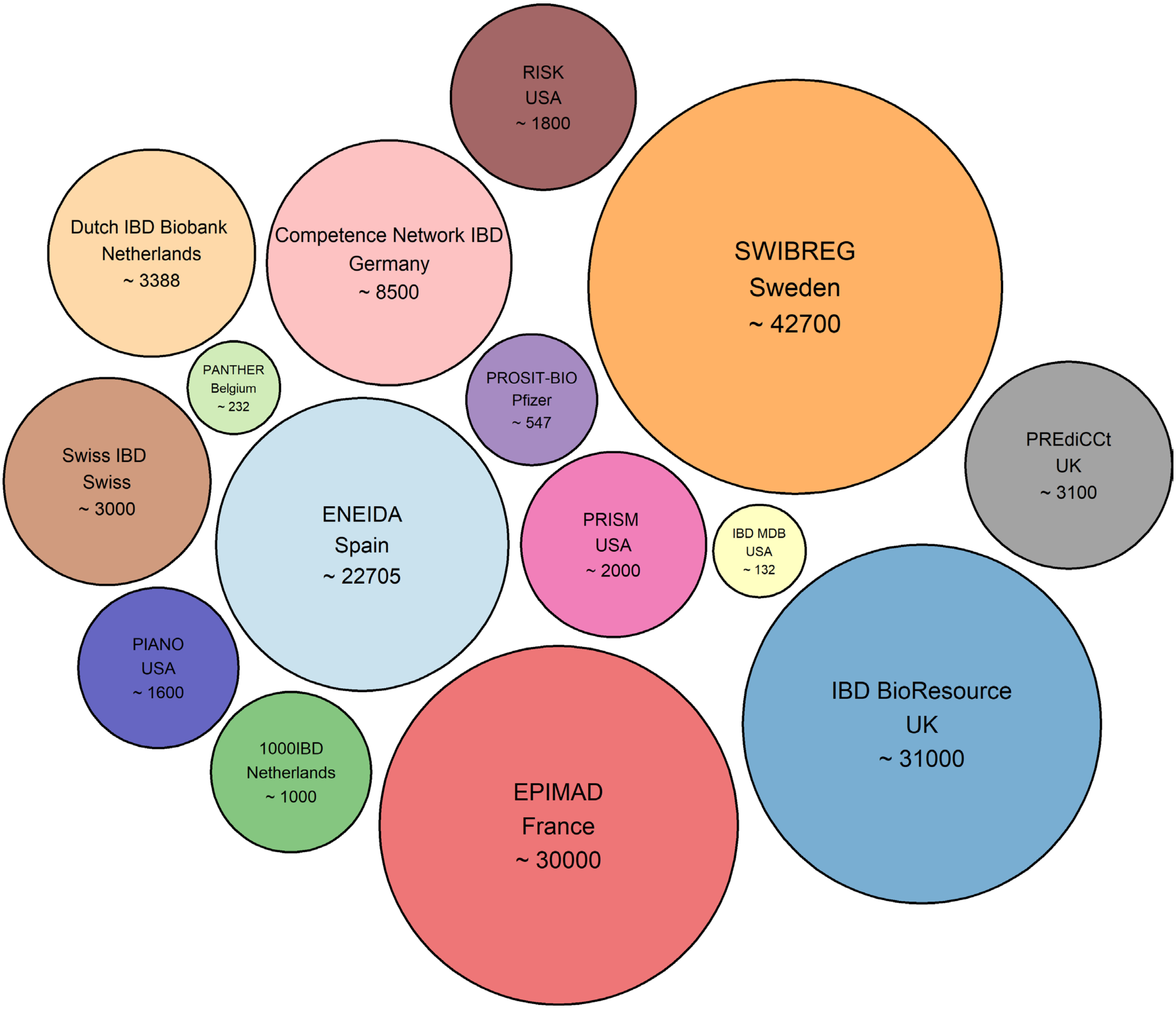

Figure 3 Academic initiatives with cohorts/biobanks in IBD. The numbers in each circle represent the approximate patient cohort size.

degree through extraction of the intensity and distribution of red pixels in the red density score in UC.[87 88] Similarly, assessment of CT/MRI images in IBD is extremely subjective; therefore, computer-aided scores could potentially overcome interobserver variation. A semiautomated image analysis software showed a performance similar to those of experienced radiologists for the assessment of CD structural bowel damage in abdominal CT-enterography data.[89] Also, machine learning methods and algorithms have been applied to predict the grading of severity of CD in abdominal MRI data.[90 91] Additionally, machine learning algorithms could assist with the time-consuming assessment of wireless capsule endoscopy data. It paves the way for automated analysis of wireless capsule endoscopy images to detect CD lesions via detection of predefined structural and textural characteristics, as well as enhancement of the underlying pixel information.[92 93]

Machine learning may also improve the analysis of histopathology and possibly tackle the unmet need of patients with unclassified IBD. Raman microspectroscopy as a cell and tissue diagnostics approach has been investigated to distinguish different IBD subtypes. Bielecki *et al* proposed that a machine learning-based workflow is capable of distinguishing tissue morphology among healthy subjects, CD and UC with great accuracy.[94] Ultimately, artificial intelligence is promising in medical imaging and will undoubtedly have a considerable impact on endoscopy practice in the future (figure 4).

Predicting prognosis

Predicting disease progression and severity is pivotal to the design of appropriate disease management strategies for individual patients. Machine learning has the potential to assist with this. Extraction of information from routinely collected electronic medical records (EMRs), such as physician's clinical observations and endoscopy reports, will allow researchers to perform prognostic research on longitudinal data. A machine learning model trained on codified information (International Classification of Diseases, Ninth Revision (ICD-9)) retrieved from EMRs, including a set of baseline laboratory parameters, patient demographics and clinical characteristics, accurately (AUC=0.93) predicted disease severity in patients with CD.[95] Similarly, Waljee *et al* constructed a random forests machine learning model to

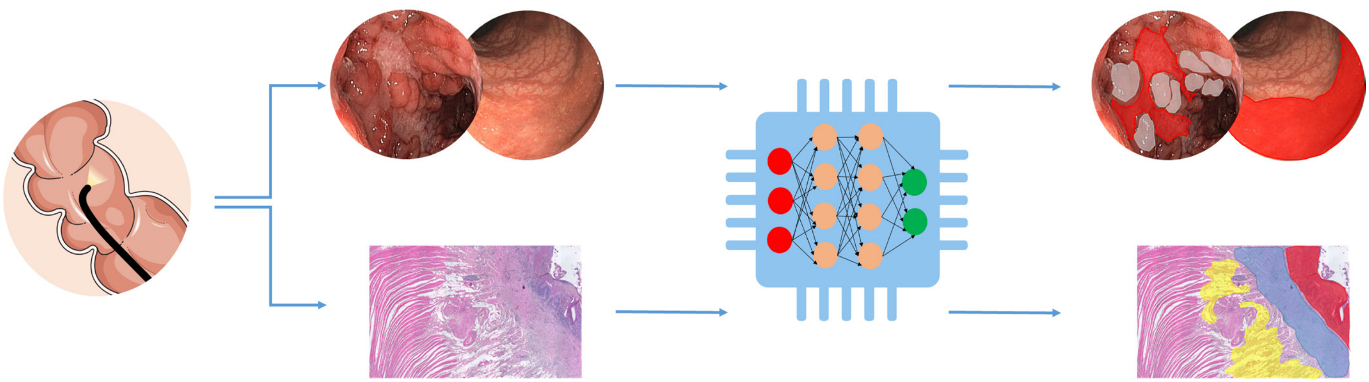

Figure 4 Artificial intelligence in medical imaging. Graphical representation of a simple deep learning-based image segmentation approach to predict boundaries of inflamed areas. The top section of the figure represents the endoscopic image of colonic CD demonstrating the 'cobblestone' appearance and ulceration. Using a simple deep learning-based image segmentation method inflamed boundaries could be predicted: cobblestone in grey and inflamed ulcer in red. The bottom section of the figure illustrates a histopathology image of inflamed stenosis from ileal CD. A deep learning-based method could be used for image segmentation and predicting boundaries of inflamed areas: acute infiltration (ulcer) in red, muscolari mucosae thickening in blue and adipocytes hyperplasia in yellow. CD, Crohn's disease.

predict IBD-related hospitalisation and outpatient steroid use, as surrogate markers of disease flares (AUC=0.87, 95% CI 0.87 to 0.88). The authors pointed out older age, high serum albumin, platelet counts, immunosuppressive medication, history of corticosteroid use and hospitalisation as risk predictors.[96] One way to improve and facilitate data extraction from plain text in medical records is by employing natural language processing (NLP), another field of artificial intelligence.[97] For example, an NLP-based model showed superior performance in comparison to an ICD-9-based model for extracting extraintestinal manifestation data from EMRs.[98] In the IBD therapeutic space, Cai *et al* applied NLP to clinical notes in identifying the risk of arthralgia in two groups of IBD patients: one treated with vedolizumab and another with TNF inhibitor.[98] Hou *et al* examined the performance of NLP-based software to classify the endoscopy procedure in patients with IBD that was performed in a diagnostic or follow-up context by mining the pathology reports.[99]

Most investigation in prognostic research has centred on investigating the diversity of underlying disease pathophysiology aiming to identify predictive correlates, which shed light onto the factors prompting disease progress, severity and clinical manifestations. Genome-wide association studies in patients with CD pinpointed the distinct genetic bases of susceptibility and prognosis and hence separate biology. These prognosis-associated SNPs are enriched for pathways involved in the regulation of innate and adaptive immune responses and responses to microorganisms. Among those, four loci have been identified to be significantly associated with prognosis in CD, namely, *FOXO3*, *XACT*, a region upstream of *IGFBP1* and *MHC*. This serves as the point of departure for better understanding of the biology that determines disease prognosis.[100] Advancement of omics techniques and data analytics have led to molecular and functional-based disease classification. For example, on combining mucosal gene expression, metagenomics and CD4 +T cell population signatures, Tang *et al* employed a machine learning approach to define a list of 26 predictors, which were effective in distinguishing between normal intestinal regions and those with active inflammation in IBD patients. Using network analysis to further interpret the inferred predictors, the authors pinpointed the role of *SAA1* in the induction of IL17 and IL22 secretion by CD4+ T cells in relation to *Bacteroides* abundance.[101]

To date, various studies have assessed the predictive value of gut microbiota. Machine learning models, especially random forests, are used extensively in microbiome research due to their ease of understanding, excellent performance and incorporated feature selection (via estimating feature importance). Douglas and colleagues studied microbial taxa and their inferred function in intestinal biopsies of 20 treatment-naive paediatric patients with CD and 20 control patients. The authors pointed to the predictive value of microbiome profiling using 16 s rRNA sequencing for the disease state, whereas metagenomic-based identified markers performed best for classifying treatment response.[102]

When large integrated EMRs and multiomics datasets are combined with a powerful and robust machine learning framework, they can achieve exceptional results. Cushing *et al* identified a unique expression profile in anti-TNF-naive and anti-TNF-exposed patients with CD that could predict postoperative disease recurrence. The authors uncovered 30 influential transcripts in anti-TNF-naive patients using random forests-based machine learning models built on demographic and clinical data extracted from the EMR and transcriptomic profile of non-inflamed ileal tissue.[103]

These methodologies provide a promising initiative to the application of machine learning to predict IBD disease course and outcome, a research scope demanding comprehensive and longitudinal investigations. By expansion of data resources as well as advancement in analytic approaches, prediction of prognosis and identifying low-risk and high-risk patients doubtlessly become feasible. Future studies should aim at mining health records and integrating them with multiomics data.

Predicting drug response

In the past decades, enormous efforts have been made to predict the response to medications. Since prospective indicators of drug responses are expected to have a big impact on pharmacoeconomics, machine learning approaches have been applied to dissect the underlying complexities and predict responses to drugs used in IBD treatments. Integration of clinical and laboratory data has been used for monitoring drugs with narrow therapeutic window, such as thiopurine, to assess the risk of developing adverse events. Currently, evaluation of clinical efficacy and risk management of thiopurine is either through blood count or measuring and monitoring of the level of its metabolites 6-thioguaninenucleotide, as an indicator of response, and

6-methylmercaptopurine, which is associated with the risk of hepatotoxicity. Waljee and colleagues studied the predictive value of a set of clinical and laboratory data to differentiate clinical responders from non-responders using a machine learning model, random forests. The proposed model has an AUC of 0.85, in contrast to the conventional model with an AUC of 0.59.[104] Subsequent work has shown significant clinical benefits, including decreased steroid prescriptions, hospitalisations and surgeries.[105]

Using clinical trial data from the GEMINI I and GEMINI II studies with vedolizumab, Waljee and colleagues developed a machine learning model, random forests, incorporating demographic data, clinical data and laboratory tests to predict the likelihood of achieving week 52 corticosteroid-free endoscopic remission in patients with UC[106] and CD[107] treated with vedolizumab. Interestingly, the strongest positive prognostic markers in patients with UC were low levels of faecal calprotectin and albumin; and those in patients with CD were low levels of serum C reactive protein and albumin.

An example of efforts to generate and integrate molecular and clinical data to guide treatment relates to identifying biomarkers predictive of drug response. In an interesting study, Zarringhalam and colleagues searched for predictive biomarkers for response to infliximab for refractory UC. First, an in-house algorithm incorporating causal prior knowledge (relationships between genes defined from the literature) with gene expression data was used to define upstream gene regulators. The newly defined features were subsequently used in a machine learning model (panelised logistic regression) to predict patient's response to infliximab (accuracy=70%). The authors pinpointed interferon gamma (IFNG), lipopolysaccharide (LPS) and TNF as key regulators. They inferred that the lack of response could be due to higher expression of the TNF pathway components, enzymatic dysregulation in the IFNG pathway and activation of the LPS–TLR4 pathway triggered by the presence of Gram-negative bacteria.[108]

Given that the human gut hosts billions of microorganisms, the gut microbiome is increasingly known to be a contributor of drug efficacy.[109] Doherty and colleagues used a machine learning model using the concept of random forests to predict the therapeutic response to ustekinumab in patients with CD.[110] The model helped in the identification of microbial signatures such as altered levels of *Faecalibacterium* that were predictive of remission. Similarly, Shaw *et al* performed an analysis using a similar classifier model based on longitudinal microbiome data derived from 19 treatment-naive paediatric individuals diagnosed with IBD and exposed to biologics.[111] The authors were able to achieve a 76.5% accuracy in predicting responders based on the pretreatment microbiome. These studies suggest that stratification of patients according to their molecular and clinical characteristics would be beneficial for evaluating therapeutic efficacy. Multiomics data integration could prove useful in biomarker discovery for treatment response. Recently, our group identified 10-feature transcriptomic (accuracy of 98%) and 15-feature genomic (accuracy 96.6%) panels predicting endoscopic response to ustekinumab by incorporating genomics and transcriptomics data into a matrix factorisation-based machine learning model in patients with CD.[112]

KEY CHALLENGES AND OPPORTUNITIES

Big data and artificial intelligence represent a great step forward in precision medicine with a high reward stand-off. With the potential to simultaneously discover new therapies, make informed treatment decisions and identify disease subgroups, there is a massive effort towards making artificial intelligence commonplace in clinical and biomedical research. The increasing availability of big data, especially multiomics datasets from large IBD cohorts, development of machine learning-based algorithms and systems biology-based tools have enabled the discovery of biological knowledge relevant to IBD. However, key challenges remain especially in the realm of how such datasets become useful in clinical translation and precision medicine (figure 5). Even though existing datasets have yielded interesting biological insights, the number of cases of such datasets resulting in direct clinical benefits, has been few and far in between. This is striking especially given the fact that there is a call for personalised therapies.

This translational gap is not unsurprising since the causality axis for IBD has not yet been established. In part, this could be attributed to the temporal nature (cross sectional or longitudinal) and/or the composition (type of multiomics data types) of datasets. Longitudinal profiling of multiomics datasets even from smaller cohorts may have higher performance and information richness than larger cohorts without longitudinal profiling. This has been demonstrated in other complex diseases such as diabetes and obesity.[113 114] The cross-sectional nature of most IBD datasets tends to limit their usefulness in inferring causal mechanisms.

Missing data are also a key challenge since these leave researchers with a choice of having to leave out particular samples or imputing missing data points, which results in reduced data and unintended errors, respectively. Also unbalanced distribution of clinical or phenotypical heterogeneity is a real-world issue affecting the interpretation of any integrative analyses. There is also a dearth of omics datasets such as proteomics, which are closer to phenotypical manifestations than other data types such as genotyping or transcriptomics. The availability of already assembled large IBD cohorts with stored biomaterial throws open multiple opportunities for improving and delivering on the research front. Sampling the biomaterials for generating the missing datatypes provides new opportunities to explore complete datasets. Thus, coordination between lead researchers and funding agencies to generate coherent multilayered datasets from the same patient samples is a major requirement. Harmonised collection, storage and usage of patient metadata and medical records are also a key challenge for inferring knowledge and clinical translation.

The contribution of disease complexity to the usefulness of multiomics datasets also extends to the composition and completeness of these datasets. The specific roles of distinct cellular populations and lineages in driving and contributing to specific phenotypes are becoming increasingly clear in IBD.[115–118] Adding to the complexity is the recently discovered fact that mutations occur in a cell type-specific manner.[119] Most of the datasets from organised cohorts have either profiled expression and genotyping from bulk RNA and DNA extracted from biopsy material or whole blood respectively, making it difficult to investigate the role of specific cell types in the aetiology and pathogenesis of IBD. As a case in point, Smillie *et al* demonstrated the power of profiling the expression of more than 50 cell types to pinpoint intercellular circuits which distinguish UC and healthy states.[120]

The implementation of big data and artificial intelligence approaches into clinical practice and meaningful benefits for patients is the ultimate challenge. On one hand, the deployment and operationalisation of big data are challenging, which are being addressed using computational sciences and algorithmic

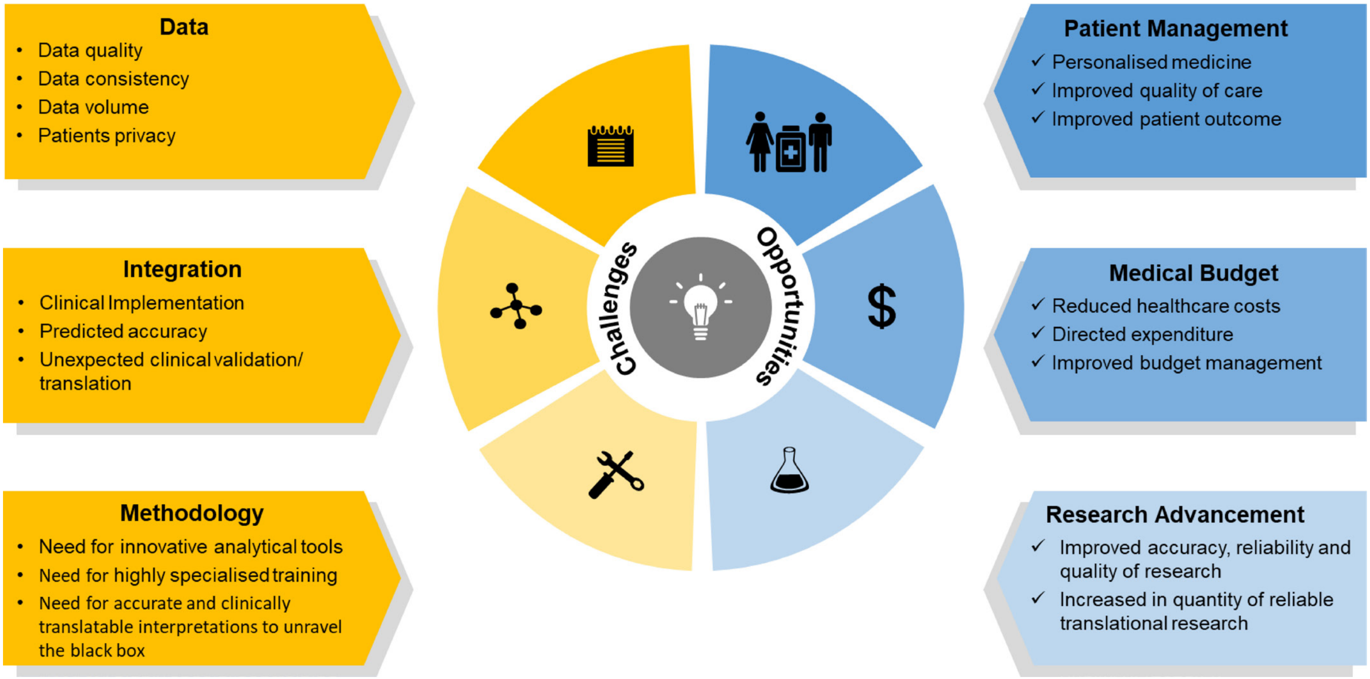
Data
Data quality
Data consistency
Data volume
Patients privacy
Integration
Clinical Implementation
Predicted accuracy
Unexpected clinical validation/ translation
Methodology
Need for innovative analytical tools
Need for highly specialised training
Need for accurate and clinically translatable interpretations to unravel the black box
Challenges
Opportunities
$
Patient Management
✓ Personalised medicine
✓ Improved quality of care
✓ Improved patient outcome
Medical Budget
✓ Reduced healthcare costs
✓ Directed expenditure
✓ Improved budget management
Research Advancement
✓ Improved accuracy, reliability and quality of research
✓ Increased in quantity of reliable translational research

Figure 5 Opportunities and challenges in the use of machine learning and data integration to achieve improved and personalised healthcare in IBD. While challenges exist in generating good quality data in a standardised manner and at a volume deemed suitable for ensuring baseline performance of machine learning models, there remain difficulties in terms of the expertise needed to identify and employ appropriate tools for data integration and interpretation. However, with emerging advances in the data integration field, the incentives and opportunities to advance precision medicine with clinical implications are expected to drive integrative IBD research forward.

frameworks to manage problems related to storage, analysis, integration and interpretation of big data. Most of the infrastructures are being explored and adopted from the computer science field into healthcare. These include cloud-based data storage and analysis, and massively parallel processing hardware to tackle the rapid increase in the volumes of data from EMR, imaging and omics measurements, for example. Moreover, there is a need for user-friendly software and workflows to facilitate the integration of big data analytics into clinical practice. For instance, there have been efforts into developing NLP-based software to assist medical investigators with extracting data from plain text, such as clinical reports.[121 122]

On the other hand, many clinicians are cautious of artificial intelligence approaches mainly because most of these approaches are essentially black boxes and do not link predictions to underlying mechanisms, nor provide functional explanations for the discovered associations, correlations and recommended decisions. However, causal mechanistic insights are key for clinical applicability so as to enhance reliability and thereby patient safety, especially in a complex heterogeneous disease such as IBD. Furthermore, as poorly validated models could do more harm than good, in depth experimental and clinical validation is crucial for machine learning-based models before implementation in clinical setting. From the analytics point, interpretable machine learning models should be developed.[123] Besides, there is a need to benchmark performance indices and parameters to evaluate the performance of machine learning techniques.[124] Other challenges include the uncertainties associated with analyses involving the use of biological networks despite the functional context provided by the networks. Even though high-quality manually curated and benchmarked networks exist,[125 126] analytical methods which take into account the uncertainties of individual interactions and their contextuality need to be developed. Clinical validation is fundamental for the implementation of artificial intelligence-based approaches. In one of the first randomised clinical trials using artificial intelligence, Lin *et al* compared the efficacies of childhood cataracts diagnosed by senior ophthalmologists with those from CC-Cruiser, a previously developed artificial intelligence platform for risk stratification and treatment guidance. This trial showed that regardless of the inferior accuracy of CC-Cruiser compared with senior ophthalmologists, artificial intelligence had the capacity to assist doctors in decision-making.[127 128] All in all, clinicians are right to be sceptical of the implementation of these otherwise inexplicable approaches in clinical practice, and although there have been considerable advances in the implementation of big data, there still remain many technological, translational and cultural barriers for the assimilation of artificial intelligence approaches into clinical practice.

CONCLUSION

By enabling data integration and assisting the discovery of nontrivial patterns and translatable knowledge in the integrated datasets, machine learning and systems biology offer unique opportunities to study and investigate the aetiology of complex diseases such as IBD. Machine learning guided IBD research has great potential to accelerate the formulation of cutting-edge precision medicine applications with clinical relevance and utility. However, for the promise of machine learning to come to translational fruition, there remain many stumbling blocks. However, almost all of the challenges also come with a huge potential for discovering knowledge and translating it to IBD clinical practice. It is expected that, with the availability of large IBD initiatives such as national biobanks with stored biomaterial, datasets can be made more coherent and complete, thus filling the biological gap for systems biology and the statistical gap for machine learning to produce knowledge which is closer to clinical practice and translation.

SEARCH STRATEGY

Articles were retrieved from PubMed after employing the following search criteria. Two key-word groups were created, with the first one comprising "Inflammatory Bowel Disease", "Crohn's disease" and "ulcerative colitis" and second one comprising "machine learning", "Artificial Intelligence", "deep learning", "-omics", "big data", "systems biology", "network biology", "genomics", "transcriptomics", "GWAS", "proteomics" and "microbiome". Pairwise combination of keywords from the two groups was used to search for articles published until July 2019. Only articles written in English were included.

Contributors NSST performed the literature search, wrote the individual sections, compiled the visual objects and was involved with the overall manuscript generation. PS wrote an individual section of the manuscript and made a critical review of the manuscript. MM performed the literature survey and critical review of the manuscript. TK performed a critical review of the manuscript. BV performed a critical review of the manuscript. SV formulated the idea and read and edited the manuscript. All authors discussed and revised the draft and approved the final version of the manuscript.

Funding This work was supported by European Research Council (ERC). NSST and PS were supported by the ERC Advanced Grant (ERC-2015-AdG, 694679, CrUCCial). BV is a doctoral fellow and SV is a senior clinical investigator of the Research Foundation Flanders (FWO), Belgium. MM is supported by Biotechnological and Biosciences Research Council (BBSRC) Norwich Research Park Biosciences Doctoral Training Partnership (grant number BB/S50743X/1), as an NPIF Award. TK was supported by a fellowship in computational biology at the Earlham Institute (Norwich, UK) in partnership with the Quadram Institute (Norwich, UK) and was strategically supported by the BBSRC (BB/J004529/1, BB/P016774/1 and BB/CSP17270/1).

Competing interests BV received lecture fees from Abbvie, Ferring Pharmaceuticals, Janssen, R-Biopharm and Takeda; consultancy fees from Janssen and Sandoz. SV: research grant: MSD, AbbVie, Takeda, Pfizer, J&J; lecture fee: MSD, AbbVie, Takeda, Ferring, Centocor, Hospira, Pfizer, J&J, Genentech/Roche; consultancy: MSD, AbbVie, Takeda, Ferring, Centocor, Hospira, Pfizer, J&J, Genentech/Roche, Celgene, Mundipharma, Celltrion, SecondGenome, Prometheus, Shire, Prodigest, Gilead, Galapagos. SV is a senior clinical investigator of the Research Foundation–Flanders (FWO). The work of MM and TK is supported by BenevolentAI, and TK's work is also supported by Unilever.

Patient consent for publication Not required.

Provenance and peer review Commissioned; externally peer reviewed.

ORCID iDs

Nasim Sadat Seyed Tabib http://orcid.org/0000-0001-9612-0012
Bram Verstockt http://orcid.org/0000-0003-3898-7093

REFERENCES

1 Korcsmaros T, Schneider MV, Superti-Furga G. Next generation of network medicine: interdisciplinary signaling approaches. *Integr Biol* 2017;9:97–108.

2 Weersma RK, Xavier RJ, Vermeire S, *et al*. Multiomics analyses to deliver the most effective treatment to every patient with inflammatory bowel disease. *Gastroenterology* 2018;155:e1–4.

3 Ananthakrishnan AN. Epidemiology and risk factors for IBD. *Nat Rev Gastroenterol Hepatol* 2015;12:205–17.

4 de Souza HSP, Fiocchi C. Immunopathogenesis of IBD: current state of the art. *Nat Rev Gastroenterol Hepatol* 2016;13:13–27.

5 Ananthakrishnan AN, Bernstein CN, Iliopoulos D, *et al*. Environmental triggers in IBD: a review of progress and evidence. *Nat Rev Gastroenterol Hepatol* 2018;15:39–49.

6 Gecse KB, Vermeire S. Differential diagnosis of inflammatory bowel disease: imitations and complications. *Lancet Gastroenterol Hepatol* 2018;3:644–53.

7 Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Health Inf Sci Syst* 2014;2:3.

8 Brooks J, Watson A, Korcsmaros T. Omics approaches to identify potential biomarkers of inflammatory diseases in the focal adhesion complex. *Genomics Proteomics Bioinformatics* 2017;15:101–9.

9 Gligorijević V, Pržulj N. Methods for biological data integration: perspectives and challenges. *J R Soc Interface* 2015;12:20150571.

10 Ideker T, Galitski T, Hood L. A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* 2001;2:343–72.

11 Camacho DM, Collins KM, Powers RK, *et al*. Next-Generation machine learning for biological networks. *Cell* 2018;173:1581–92.

12 Hood L, Tian Q. Systems approaches to biology and disease enable translational systems medicine. *Genomics Proteomics Bioinformatics* 2012;10:181–5.

13 Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol* 2019;20:e262–73.

14 Brooks J, Modos D, Sudhakar P, *et al*. A systems genomics approach to uncover patient-specific pathogenic pathways and proteins in a complex disease. *bioRxiv* 2019;692269.

15 Sorzano COS, Vargas J, Pascual-Montano AD. A survey of dimensionality reduction techniques. *ArXiv* 2014:abs/1403.2.

16 Tedjo DI, Smolinska A, Savelkoul PH, *et al*. The fecal microbiota as a biomarker for disease activity in Crohn's disease. *Sci Rep* 2016;6:35216.

17 Mirkov MU, Verstockt B, Cleynen I. Genetics of inflammatory bowel disease: beyond NOD2. *Lancet Gastroenterol Hepatol* 2017;2:224–34.

18 Ogura Y, Bonen DK, Inohara N, *et al*. A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature* 2001;411:603–6.

19 Shah SC, Kusiak A. Data mining and genetic algorithm based gene/SNP selection. *Artif Intell Med* 2004;31:183–96.

20 Phuong TM, Lin Z, Altman RB. Choosing SNPs using feature selection. In: 2005 IEEE Computational Systems Bioinformatics Conference (CSB'05). *IEEE* 2005:301–9.

21 Long N, Gianola D, Rosa GJM, *et al*. Machine learning classification procedure for selecting SNPs in genomic selection: application to early mortality in broilers. *J Anim Breed Genet* 2007;124:377–89.

22 Bermingham ML, Pong-Wong R, Spiliopoulou A, *et al*. Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Sci Rep* 2015;5:10312.

23 Wei Z, Wang W, Bradfield J, *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;92:1008–12.

24 Luo Y, de Lange KM, Jostins L, *et al*. Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nat Genet* 2017;49:186–92.

25 Romagnoni A, Jégou S, Van Steen K, *et al*. Comparative performances of machine learning methods for classifying Crohn disease patients using genome-wide genotyping data. *Sci Rep* 2019;9:10351.

26 Schwerd T, Bryant RV, Pandey S, *et al*. Nox1 loss-of-function genetic variants in patients with inflammatory bowel disease. *Mucosal Immunol* 2018;11:562–74.

27 Ghandi M, Lee D, Mohammad-Noori M, *et al*. Enhanced regulatory sequence prediction using gapped k-mer features. *PLoS Comput Biol* 2014;10:e1003711.

28 Alipanahi B, Delong A, Weirauch MT, *et al*. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 2015;33:831–8.

29 Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. *Nat Methods* 2015;12:931–4.

30 Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res* 2016;26:990–9.

31 Zou J, Huss M, Abid A, *et al*. A primer on deep learning in genomics. *Nat Genet* 2019;51:12–18.

32 Li XL, Zhou CY, Sun Y, *et al*. Bioinformatic analysis of potential candidates for therapy of inflammatory bowel disease. *Eur Rev Med Pharmacol Sci* 2015;19:4275–84.

33 Hong SN, Joung J-G, Bae JS, *et al*. Rna-Seq reveals transcriptomic differences in inflamed and Noninflamed intestinal mucosa of Crohn's disease patients compared with normal mucosa of healthy controls. *Inflamm Bowel Dis* 2017;23:1098–108.

34 Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol* 2005;4:Article17.

35 Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* 2008;9:559.

36 Lin X, Li J, Zhao Q, *et al*. WGCNA reveals key roles of IL8 and MMP-9 in progression of involvement area in colon of patients with ulcerative colitis. *Curr Med Sci* 2018;38:252–8.

37 Verstockt S, De Hertogh G, Van der Goten J, *et al*. Gene and Mirna Regulatory Networks During Different Stages of Crohn's Disease. *J Crohn's Colitis* 2019;13:916–30.

38 Verstockt B, Verstockt S, Creyns B, *et al*. Mucosal IL13RA2 expression predicts nonresponse to anti-TNF therapy in Crohn's disease. *Aliment Pharmacol Ther* 2019;49:572–81.

39 Thiele I, Swainston N, Fleming RMT, *et al*. A community-driven global reconstruction of human metabolism. *Nat Biotechnol* 2013;31:419–25.

40 Knecht C, Fretter C, Rosenstiel P, *et al*. Distinct metabolic network states manifest in the gene expression profiles of pediatric inflammatory bowel disease patients and controls. *Sci Rep* 2016;6:32584.

41 Yuan F, Zhang Y-H, Kong X-Y, *et al*. Identification of candidate genes related to inflammatory bowel disease using minimum redundancy maximum relevance, incremental feature selection, and the shortest-path approach. *Biomed Res Int* 2017;2017:1–15.
42 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;23:1516–23.
43 Manichanh C, Borruel N, Casellas F, *et al*. The gut microbiota in IBD. *Nat Rev Gastroenterol Hepatol* 2012;9:599–608.
44 Halfvarson J, Brislawn CJ, Lamendella R, *et al*. Dynamics of the human gut microbiome in inflammatory bowel disease. *Nat Microbiol* 2017;2:17004.
45 Norman JM, Handley SA, Baldridge MT, *et al*. Disease-Specific alterations in the enteric Virome in inflammatory bowel disease. *Cell* 2015;160:447–60.
46 Zuo T, Lu X-J, Zhang Y, *et al*. Gut mucosal virome alterations in ulcerative colitis. *Gut* 2019;68:1169–79.
47 Greenblum S, Turnbaugh PJ, Borenstein E. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. *Proc Natl Acad Sci U S A* 2012;109:594–9.
48 Li X, LeBlanc J, Elashoff D, *et al*. Microgeographic proteomic networks of the human colonic mucosa and their association with inflammatory bowel disease. *Cell Mol Gastroenterol Hepatol* 2016;2:567–83.
49 Erickson AR, Cantarel BL, Lamendella R, *et al*. Integrated Metagenomics/ Metaproteomics reveals human Host-Microbiota signatures of Crohn's disease. *PLoS One* 2012;7:e49138.
50 Häsler R, Sheibani-Tezerji R, Sinha A, *et al*. Uncoupling of mucosal gene regulation, mRNA splicing and adherent microbiota signatures in inflammatory bowel disease. *Gut* 2017;66:2087–97.
51 Lloyd-Price J, Arze C, Ananthakrishnan AN, *et al*. Multi-Omics of the gut microbial ecosystem in inflammatory bowel diseases. *Nature* 2019;569:655–62.
52 Verstockt B, Ferrante M, Vermeire S, *et al*. New treatment options for inflammatory bowel diseases. *J Gastroenterol* 2018;53:585–90.
53 Maaser C, Sturm A, Vavricka SR, *et al*. ECCO-ESGAR guideline for diagnostic assessment in IBD Part 1: initial diagnosis, monitoring of known IBD, detection of complications. *J Crohn's Colitis* 2019;13:144–64.
54 Sturm A, Maaser C, Calabrese E, *et al*. ECCO-ESGAR guideline for diagnostic assessment in IBD Part 2: IBD scores and general principles and technical aspects. *J Crohn's Colitis* 2019;13:273–84.
55 Colombel J-F, Panaccione R, Bossuyt P, *et al*. Effect of tight control management on Crohn's disease (calm): a multicentre, randomised, controlled phase 3 trial. *Lancet* 2017;390:2779–89.
56 Bremer V, Becker D, Kolovos S, *et al*. Predicting therapy success and costs for personalized treatment recommendations using baseline characteristics: data-driven analysis. *J Med Internet Res* 2018;20:e10275.
57 Wang H-Y, Hung C-C, Chen C-H, *et al*. Increase Trichomonas vaginalis detection based on urine routine analysis through a machine learning approach. *Sci Rep* 2019;9.
58 Cleynen I, Boucher G, Jostins L, *et al*. Inherited determinants of Crohn's disease and ulcerative colitis phenotypes: a genetic association study. *Lancet* 2016;387:156–67.
59 Jeong C-S, Kim D. Inferring Crohn's disease association from exome sequences by integrating biological knowledge. *BMC Med Genomics* 2016;9:35.
60 Medina I, Montaner D, Tarraga J, *et al*. Prophet, a web-based tool for class prediction using microarray data. *Bioinformatics* 2007;23:390–1.
61 Montero-Meléndez T, Llor X, García-Planella E, *et al*. Identification of novel predictor classifiers for inflammatory bowel disease by gene expression profiling. *PLoS One* 2013;8:e76235.
62 Han L, Maciejewski M, Brockel C, *et al*. A probabilistic pathway score (PROPS) for classification with applications to inflammatory bowel disease. *Bioinformatics* 2018;34:985–93.
63 Seeley EH, Washington MK, Caprioli RM, *et al*. Proteomic patterns of colonic mucosal tissues delineate Crohn's colitis and ulcerative colitis. *Proteomics Clin Appl* 2013;7:541–9.
64 Cao B, Zhou X, Ma J, *et al*. Role of miRNAs in inflammatory bowel disease. *Dig Dis Sci* 2017;62:1426–38.
65 Duttagupta R, DiRienzo S, Jiang R, *et al*. Genome-Wide maps of circulating miRNA biomarkers for ulcerative colitis. *PLoS One* 2012;7:e31241.
66 Hübenthal M, Hemmrich-Stanisak G, Degenhardt F, *et al*. Sparse modeling reveals miRNA signatures for diagnostics of inflammatory bowel disease. *PLoS One* 2015;10:e0140155.
67 Plevy S, Silverberg MS, Lockton S, *et al*. Combined serological, genetic, and inflammatory markers differentiate Non-IBD, Crohn's disease, and ulcerative colitis patients. *Inflamm Bowel Dis* 2013;19:1139–48.
68 Morgan XC, Tickle TL, Sokol H, *et al*. Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. *Genome Biol* 2012;13:R79.
69 Pasolli E, Truong DT, Malik F, *et al*. Machine learning meta-analysis of large metagenomic datasets: tools and biological insights. *PLoS Comput Biol* 2016;12:e1004977.
70 Fioravanti D, Giarratano Y, Maggio V, *et al*. Phylogenetic convolutional neural networks in metagenomics. *BMC Bioinformatics* 2018;19:49.
71 Gevers D, Kugathasan S, Denson LA, *et al*. The treatment-naive microbiome in new-onset Crohn's disease. *Cell Host Microbe* 2014;15:382–92.
72 | Pfizer for Professionals. PROSIT-BIO. Available: https://www.pfizerpro.co.uk/product/inflectra/ulcerative-colitis/support/prosit-bio-0 [Accessed 23 Aug 2019].
73 1000 IBD. Available: https://1000ibd.org/ [Accessed 23 Aug 2019].
74 Spekhorst LM, Imhann F, Festen EAM, *et al*. Cohort profile: design and first results of the Dutch IBD Biobank: a prospective, nationwide Biobank of patients with inflammatory bowel disease. *BMJ Open* 2017;7:e016695.
75 Chaparro M, Ramas M, Benítez JM, *et al*. Extracolonic cancer in inflammatory bowel disease: data from the GETECCU Eneida registry. *Am J Gastroenterol* 2017;112:1135–43.
76 Beaulieu DB, Ananthakrishnan AN, Martin C, *et al*. Use of biologic therapy by pregnant women with inflammatory bowel disease does not affect infant response to vaccines. *Clin Gastroenterol Hepatol* 2018;16:99–105.
77 IBDMDB. Home IBDMDB. Available: https://ibdmdb.org/ [Accessed 23 Aug 2019].
78 PREdiCCt. Home. Available: https://www.predicct.co.uk/ [Accessed 23 Aug 2019].
79 CSIBD PRISM registry Hospital, Boston, MA. Available: https://www.massgeneral.org/csibd/cores/clinical.aspx [Accessed 9 Jan 2020].
80 IBD BioResource. Translating today's science into tomorrow's treatments. Available: https://www.ibdbioresource.nihr.ac.uk/ [Accessed 9 Jan 2020].
81 Home - SWISS IBDcohort. Available: http://www.ibdcohort.ch/ [Accessed 9 Jan 2020].
82 Swibreg. Patient. Available: http://www.swibreg.se/ [Accessed 9 Jan 2020].
83 EPIMAD : le plus grand registre au monde – Observatoire National des MICI. Available: http://www.observatoire-crohn-rch.fr/epimad-le-plus-grand-registre-de-malades-au-monde/ [Accessed 9 Jan 2020].
84 Study management - Competence Network for Bowel Diseases. Available: http://www.kompetenznetz-darmerkrankungen.de/Studienmanagement [Accessed 9 Jan 2020].
85 Bossuyt P, Vermeire S, Bisschops R. Scoring endoscopic disease activity in IBD: artificial intelligence sees more and better than we do. *Gut* 2020;69:788–9.
86 Stidham RW, Liu W, Bishu S, *et al*. Performance of a deep learning model vs human reviewers in grading endoscopic disease severity of patients with ulcerative colitis. *JAMA Netw Open* 2019;2:e193963.
87 Bossuyt P, Nakase H, Vermeire S, *et al*. 436 - Automated Digital Calculation of Endoscopic Inflammation in Ulcerative Colitis: Results of the Red Density Study. *Gastroenterology* 2018;154:S98–9.
88 Bossuyt P, Nakase H, Vermeire S, *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. doi: 10.1136/gutjnl-2019-320056. [Epub ahead of print: 8 Jan 2020]. doi:10.1136/gutjnl-2019-320056. [Epub ahead of print:
89 Stidham RW, Enchakalody B, Waljee AK, *et al*. Assessing Small Bowel Stricturing and Morphology in Crohn's Disease Using Semi-automated Image Analysis. *Inflamm Bowel Dis*;11.
90 Tielbeek JAW, Vos FM, Stoker J. A computer-assisted model for detection of MRI signs of Crohn's disease activity: future or fiction? *Abdom Imaging* 2012;37:967–73.
91 Mahapatra D, Schüffler PJ, Tielbeek JAW, *et al*. Semi-supervised and active learning for automatic segmentation of Crohn's disease. *Med Image Comput Comput Assist Interv* 2013;16:214–21.
92 Kumar R, Qian Zhao Q, Seshamani S, *et al*. Assessment of Crohn's Disease Lesions in Wireless Capsule Endoscopy Images. *IEEE Trans Biomed Eng* 2012;59:355–62.
93 Charisis VS, Hadjileontiadis LJ. Potential of hybrid adaptive filtering in inflammatory lesion detection from capsule endoscopy images. *WJG* 2016;22:8641.
94 Bielecki C, Bocklitz TW, Schmitt M, *et al*. Classification of inflammatory bowel diseases by means of Raman spectroscopic imaging of epithelium cells. *J Biomed Opt* 2012;17:0760301.
95 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;25:1201–18.
96 Waljee AK, Lipson R, Wiitala WL, *et al*. Predicting hospitalization and outpatient corticosteroid use in inflammatory bowel disease patients using machine learning. *Inflamm Bowel Dis* 2018;24:45–53.
97 Ananthakrishnan AN, Cai T, Savova G, *et al*. Improving case definition of Crohn's disease and ulcerative colitis in electronic medical records using natural language processing. *Inflamm Bowel Dis* 2013;19:1411–20.
98 Cai T, Lin T-C, Bond A, *et al*. The association between arthralgia and vedolizumab using natural language processing. *Inflamm Bowel Dis* 2018;24:2242–6.
99 Hou JK, Chang M, Nguyen T, *et al*. Automated identification of surveillance colonoscopy in inflammatory bowel disease using natural language processing. *Dig Dis Sci* 2013;58:936–41.
100 Lee JC, Biasci D, Roberts R, *et al*. Genome-Wide association study identifies distinct genetic contributions to prognosis and susceptibility in Crohn's disease. *Nat Genet* 2017;49:262–8.
101 Tang MS, Bowcutt R, Leung JM, *et al*. Integrated analysis of biopsies from inflammatory bowel disease patients identifies SAA1 as a link between mucosal microbes with Th17 and Th22 cells. *Inflamm Bowel Dis* 2017;23:1544–54.

102 Douglas GM, Hansen R, Jones CMA, *et al*. Multi-omics differentially classify disease state and treatment outcome in pediatric Crohn's disease. *Microbiome* 2018;6:13.

103 Cushing KC, Mclean R, McDonald KG, *et al*. Predicting Risk of Postoperative Disease Recurrence in Crohn's Disease: Patients With Indolent Crohn's Disease Have Distinct Whole Transcriptome Profiles at the Time of First Surgery. *Inflamm Bowel Dis* 2019;25:180–93.

104 Waljee AK, Joyce JC, Wang S, *et al*. Algorithms Outperform metabolite tests in predicting response of patients with inflammatory bowel disease to thiopurines. *Clin Gastroenterol Hepatol* 2010;8:143–50.

105 Waljee AK, Sauder K, Patel A, *et al*. Machine learning algorithms for objective remission and clinical outcomes with thiopurines. *J Crohns Colitis* 2017;11:801–10.

106 Waljee AK, Liu B, Sauder K, *et al*. Predicting corticosteroid-free endoscopic remission with vedolizumab in ulcerative colitis. *Aliment Pharmacol Ther* 2018;47:763–72.

107 Waljee AK, Liu B, Sauder K, *et al*. Predicting Corticosteroid-Free Biologic Remission with Vedolizumab in Crohn's Disease. *Inflamm Bowel Dis* 2018;24:1185–92.

108 Zarringhalam K, Enayetallah A, Reddy P, *et al*. Robust clinical outcome prediction based on Bayesian analysis of transcriptional profiles and prior causal networks. *Bioinformatics* 2014;30:i69–77.

109 Vázquez-Baeza Y, Callewaert C, Debelius J, *et al*. Impacts of the human gut microbiome on therapeutics. *Annu Rev Pharmacol Toxicol* 2018;58:253–70.

110 Doherty MK, Ding T, Koumpouras C, *et al*. Fecal Microbiota Signatures Are Associated with Response to Ustekinumab Therapy among Crohn's Disease Patients. *MBio* 2018;9:e02120–17.

111 Shaw KA, Bertha M, Hofmekler T, *et al*. Dysbiosis, inflammation, and response to treatment: a longitudinal study of pediatric subjects with newly diagnosed inflammatory bowel disease. *Genome Med* 2016;8:75.

112 Verstockt B, Sudahakar P, Creyns B, *et al*. DOP70 An integrated multi-omics biomarker predicting endoscopic response in ustekinumab treated patients with Crohn's disease. *J Crohn's Colitis* 2019.

113 Piening BD, Zhou W, Contrepois K, *et al*. Integrative personal omics profiles during periods of weight gain and loss. *Cell Syst* 2018;6:157–70.

114 Zhou W, Sailani MR, Contrepois K, *et al*. Longitudinal multi-omics of host-microbe dynamics in prediabetes. *Nature* 2019;569:663–71.

115 Allez M, Tieng V, Nakazawa A, *et al*. CD4+NKG2D+ T cells in Crohn's disease mediate inflammatory and cytotoxic responses through MICA interactions. *Gastroenterology* 2007;132:2346–58.

116 Lee JC, Lyons PA, McKinney EF, *et al*. Gene expression profiling of CD8+ T cells predicts prognosis in patients with Crohn disease and ulcerative colitis. *J Clin Invest* 2011;121:4170–9.

117 Imam T, Park S, Kaplan MH, *et al*. Effector T helper cell subsets in inflammatory bowel diseases. *Front Immunol* 2018;9:1212.

118 Chapuy L, Bsat M, Rubio M, *et al*. IL-12 and mucosal CD14+ monocyte-like cells induce IL-8 in colonic memory CD4+ T cells of patients with Ulcerative colitis but not Crohn's disease. *J Crohn's Colitis*.

119 Yizhak K, Aguet F, Kim J, *et al*. RNA sequence analysis reveals macroscopic somatic clonal expansion across normal tissues. *Science* 2019;364:eaaw0726.

120 Smillie CS, Biton M, Ordovas-Montanes J, *et al*. Intra- and Inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* 2019;178:714–30.

121 D'Avolio LW, Nguyen TM, Farwell WR, *et al*. Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). *J Am Med Inform Assoc* 2010;17:375–82.

122 Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17:507–13.

123 Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell* 2019;1:206–15.

124 Hand DJ. Classifier technology and the illusion of progress. *Stat Sci* 2006;21:1–14.

125 Türei D, Korcsmáros T, Saez-Rodriguez J. OmniPath: guidelines and gateway for literature-curated signaling pathway resources. *Nat Methods* 2016;13:966–7.

126 Huang JK, Carlin DE, Yu MK, *et al*. Systematic evaluation of molecular networks for discovery of disease genes. *Cell Syst* 2018;6:484–95.

127 Long E, Lin H, Liu Z, *et al*. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nat Biomed Eng* 2017;1.

128 Lin H, Li R, Liu Z, *et al*. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. *EClinicalMedicine* 2019;9:52–9.