TECHNIQUES, TECHNOLOGY & ANALYSIS

Big data in IBD: a look into the future

*Pablo Olivera[1], Silvio Danese[2,3], Nicolas Jay[4], Gioacchino Natoli[3] and Laurent Peyrin-Biroulet[5]**

Abstract | Big data methodologies, made possible with the increasing generation and availability of digital data and enhanced analytical capabilities, have produced new insights to improve outcomes in many disciplines. Application of big data in the health-care sector is in its early stages, although the potential for leveraging underutilized data to gain a better understanding of disease and improve quality of care is enormous. Owing to the intrinsic characteristics of inflammatory bowel disease (IBD) and the management dilemmas that it imposes, the implementation of big data research strategies not only can complement current research efforts but also could represent the only way to disentangle the complexity of the disease. In this Review, we explore important potential applications of big data in IBD research, including predictive models of disease course and response to therapy, characterization of disease heterogeneity, drug safety and development, precision medicine and cost-effectiveness of care. We also discuss the strengths and limitations of potential data sources that big data analytics could draw from in the field of IBD, including electronic health records, clinical trial data, e-health applications and genomic, transcriptomic, proteomic, metabolomic and microbiomic data.

[1]*Gastroenterology Section, Department of Internal Medicine, Centro de Educación Médica e Investigaciones Clínicas (CEMIC), Buenos Aires, Argentina.*

[2]*IBD Center, Department of Gastroenterology, Humanitas Clinical and Research Centre, Rozzano, Milan, Italy.*

[3]*Humanitas Clinical Research Hospital, Rozzano, Milan, Italy.*

[4]*Orpailleur and Department of Medical Information, LORIA and Nancy University Hospital, Vandoeuvre-lès-Nancy, Nancy, France.*

[5]*INSERM U954 and Department of Hepatogastroenterology, Nancy University Hospital, Université de Lorraine, Vandoeuvre-lès-Nancy, Nancy, France.*

**e-mail: peyrinbiroulet@gmail.com*

https://doi.org/10.1038/s41575-019-0102-5

IBD comprises two disabling immune-mediated conditions: ulcerative colitis and Crohn's disease[1,2]. Similar to other chronic, non-infectious diseases, IBD has been classified as a prototypical complex disease[3–5], in which biological complexity arises from intricate interactions between multiple factors, such as genes, environment, microbiota and diet, among others.

During the past 20 years, major advances have been made in understanding components of IBD physiopathology, which have subsequently led to increased therapeutic options with the development of biologics and small molecule drugs engaging different targets[6]. Increases in IBD incidence and prevalence are observed worldwide but are particularly pronounced in developing countries, and this trend is expected to continue in the coming years[7,8]. This growing IBD burden will probably exacerbate current issues such as health-related costs and access to care.

Despite important breakthroughs in the past two decades, the complexity of IBD creates enormous challenges, and traditional scientific methods have been unable to address important research questions, which manifests as unmet clinical management needs[9,10]. The current paradigm of research in IBD has led to many frustrating results, and innovative methods are required to help disentangle disease complexity, which will ultimately translate into better patient care.

Theoretically, the integration of a wealth of omics data with clinical information and information on factors such as lifestyle, diet and environmental exposures could enable three major unmet clinical needs to be addressed: the identification of biomarkers that enable the early and unambiguous identification of patients with IBD before the full clinical picture has unfolded, thereby allowing very early treatment initiation; the stratification of patients by their predicted response to different drugs; and the stratification of patients by predicted disease course, which might inform the use of more or less aggressive treatment approaches.

In this Review, we explore potential applications of big data in IBD research, such as predictive models of disease course and response to therapy, characterization of disease heterogeneity, drug safety and development, precision medicine and cost-effectiveness of care. We also discuss the strengths and limitations of potential data sources that big data analytics could draw from in the field of IBD.

Big data

The increasing generation and availability of digital data in every aspect of life, coupled with enhanced analytical capability owing to advances in computational science, have produced new insights used to improve outcomes in many disciplines, notably in finance and social media. Technology giants like Google, Amazon, Facebook and Apple have successfully used big data approaches to improve sales, boost efficiency and increase earnings[11,12]. Political campaigns and government agencies have also used large data sets of information produced by citizens to develop models that guide successful electoral strategies[13].

Key points

- Big data refers to sets of data whose scale and complexity impose the use of dedicated analytical and statistical approaches.
- The distinctive attributes of big data include the four Vs: volume, variety, velocity and veracity.
- Big data approaches have been successfully used in many different areas, including finance and politics, and more recently have been increasingly implemented in health care.
- Big data analytics are innovative approaches to help disentangle the complexity of IBD.
- Potential applications of big data in the field of IBD might include precise phenomapping, the development of predictive models, precision medicine, epidemiological models and drug discovery.
- Researchers will face several potential limitations and challenges when using big data approaches in IBD, including ethical and legal restrictions, heterogeneous data sources, poor quality data and the need for validation.

Until a few years ago, the health-care sector had not substantially explored the potential benefits of big data[14]. Wide-spread implementation of big data analyses in health care is eagerly awaited because it has the potential to greatly improve many areas of care[15]. Several promising applications of big data in health care exist: better understanding of disease pathogenesis and classification of complex diseases; development of predictive prognostic models; reduction of risks; identification of predictive events to support prevention initiatives; improvement of health-care cost-effectiveness; and personalization of therapeutic regimens[16–19].

In an often-cited example of big data in health care, a paper published in 2009 reported the development of an algorithm using Google search queries to track influenza-like illnesses in the USA[20]. By monitoring and analysing the health-seeking behaviour of millions of users in the form of queries to online search engines, the appealing promise of this Google model was to predict influenza activity more rapidly than the US Centers for Disease Control and Prevention (CDC) model[20]. However, the model missed the first wave of the influenza A (H1N1) pandemic outbreak in 2009. Furthermore, it proved to be rather inaccurate: the Google model overestimated the number of medical visits for influenza-like illness by twofold compared with the CDC model[21,22]. Influenza prevalence estimates by Google are no longer published, providing an example that represents a lesson of the possible challenges ahead.

However, in a successful example published in 2016, a study explored the risk of Parkinson disease using big data methodologies by combining multiple sources of diverse data, including neuroimaging, genetic, clinical and demographic data, contained in the Parkinson disease Progression Markers Initiative archive[23]. Model-free big data machine-learning-based classification methods could predict Parkinson disease with accuracy, sensitivity and specificity consistently exceeding 96%[23].

The potential of big data in health care has been acknowledged by the US NIH. In 2013, the Big Data to Knowledge (BD2K) initiative was launched to support the research and development of innovative and transforming approaches and tools to maximize and accelerate the integration of big data and data science into biomedical research[24]. Owing to the intrinsic characteristics of IBD and the management dilemmas that it imposes, the implementation of big data research strategies not only can complement current research efforts but also could represent the only way to overcome the complexity of the disease.

Defining big data

Although an exact and universally accepted definition of big data does not exist, the concept refers to sets of data with a scale and complexity that enforces the use of dedicated analytical and statistical approaches[19,25]. In the specific case of biomedicine, big data include large-volume and high-diversity biological, genetic, clinical, environmental and lifestyle information collected from single individuals as well as large cohorts in relation to their disease and/or wellness status at one or several time points[26].

Distinctive attributes of big data include the four Vs: volume, variety, velocity and veracity[12,18,27]. The first and most obvious characteristic of big data is volume, namely, the large amount of data in a data set. Health-related data are created and accumulated continuously, and they are expected to continue to grow dramatically up to an almost unconceivable extent. The volume of health-care data was calculated at 153 exabytes in 2014, and at the projected growth rate of 48% a year, that figure is estimated to reach ~2,300 exabytes by 2020 (REFS[27,28]). This very large amount of data arises from the combination of multiple sources of structured data (for instance, administrative databases) and unstructured data (such as clinical notes), which in fact represent the second characteristic of big data: variety[17]. The third characteristic is velocity, which reflects the speed at which such information is created and accumulated. Speed is also essential to combine and analyse large and diverse data sets rapidly enough to yield valuable information to make decisions[18]. The final characteristic of big data — which is crucial in health-care informatics — is veracity[27]. Veracity means that the big data, its analytics and its outcomes provide a faithful representation of the subject under investigation as well as of the distribution of a complex phenomenon in the population. In other words, such data are expected to be unbiased and therefore intrinsically without errors and credible (although the outcome of their analyses might be affected by several technical factors)[16,18]. This characteristic is of utmost importance to reliably translate medical big data into clinical decisions. Health-care data can comprise various sources of highly variable quality, especially when considering unstructured data. Hence, veracity frequently represents a goal rather than reality[18].

Big data analytics can receive multiple inputs or data sources (FIG. 1). Theoretically, the variety of these data sources is not restrained. Currently, the most important data sources for medical big data include but are not limited to administrative databases, clinical trials registries, epidemiological studies, electronic medical records, biometric data, patient-reported health data, medical images, biomarker data, omics data (that is, genomics, proteomics and metabolomics data sets), data from social media and the internet[17,29]. The variety of potential

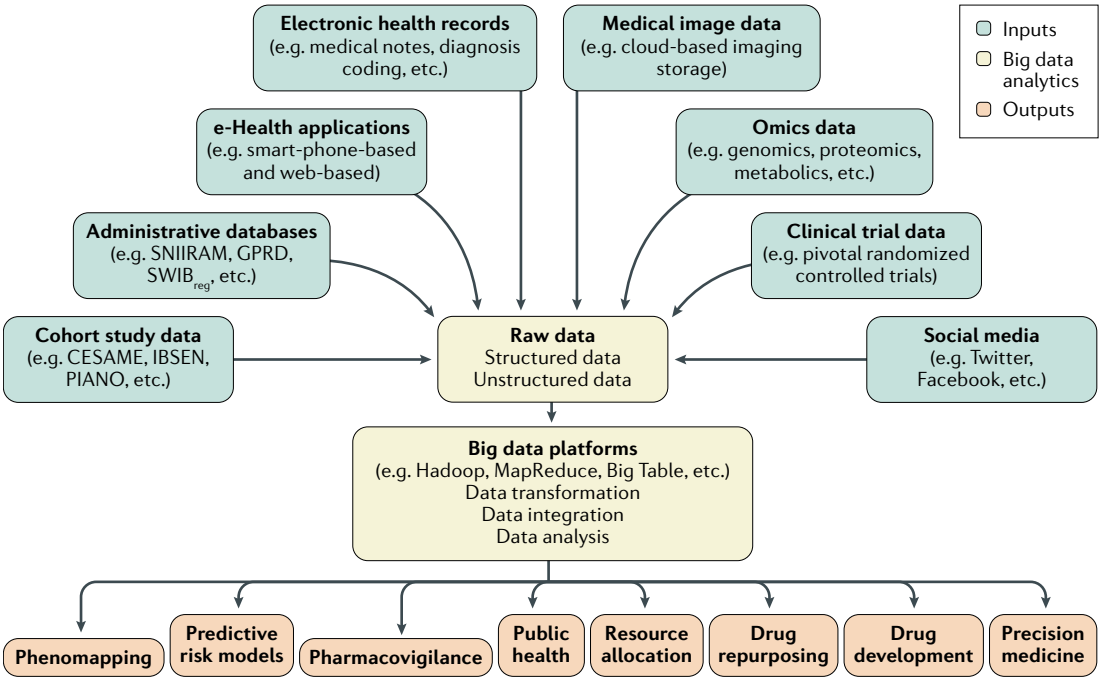

Fig. 1 | **Overview of big data in IBD.** Big data analytics in IBD research could be fed from multiple potential data sources (or inputs). Raw data from these inputs (both structured and unstructured data) need to be extracted and transformed or processed to be readily usable and stored. Big data platforms (such as Hadoop, MapReduce, Big Table, and so on) are used to organize, integrate and analyse these large volumes of data. Different analytical methods can be used, ranging from traditional statistical methods (such as regression) to advanced methods (including data mining, machine learning, clustering, text analysis and image analytics). The models developed (outputs) can then be used in different applications that might add value to current disease knowledge. CESAME, Cancer and Increased Risk Associated with Inflammatory Bowel Disease in France; GPRD, General Practice Research Database; PIANO, Pregnancy in IBD Neonatal Outcomes; SNIIRAM, Système National d'Information InterRégimes de l'Assurance Maladie; SWIB$_{reg}$, Swedish Quality Register for IBD.

data sources is expected to continue to grow, although identifying and linking together the sources that will add value and new insights represent a challenge[30].

Computer sciences have produced remarkable advances not only in hardware capacity but also in the development of software analytical platforms that enable large and diverse data sets to be handled and analysed[17]. Big data analyses use computational approaches, such as data mining and machine-learning algorithms, to extract information from a data set and to identify patterns generated by sets of features associated with disease risk, prognosis or response to therapy[11] (FIG. 1). Importantly, in most cases, these approaches return hypothesis-free predictive models, without a clear explanation of the outcome (for example, in weather forecasts, the accuracy of the prediction is important, not the complete understanding of the underlying causes). This approach contrasts with traditional hypothesis-driven scientific method research, in which hypotheses are formulated on the basis of observations, followed by design and execution of experiments and then validation of results, which ultimately leads to acceptance or rejection of the hypothesis[31]. Description of these platforms and analytical methods is beyond the scope of this Review. Analysis of health-care big data is an opportunity to discover new patterns, associations and trends that ultimately improve patient care and disease outcomes and reduce health-related costs[18].

Why do we need big data in IBD?

Disease heterogeneity. IBD has been arbitrarily divided into Crohn's disease and ulcerative colitis on the basis of descriptive characteristics, with the terms 'indeterminate colitis' or 'IBD unclassified' used when distinction is not possible[32]. IBDs are heterogeneous diseases in which a wide range of clinical phenotypes are possible, regarding not only disease location and behaviour but also age of onset, severity of symptoms, association with other immune-mediated conditions, extraintestinal manifestations, complications, response to therapy, need for surgery, and so on[33–35]. Moreover, the effect of the disease on the patient, disease burden and disease course should also be taken into account to correctly classify the disease[36]. Better classification of IBD into distinct phenotypes will not only lead to better understanding of the disease but might also help identify particular subgroups of patients that would benefit from particular interventions.

Big data approaches to disease heterogeneity might help identify these phenotypes; the hypothesis-free nature of data mining and other methodologies takes into consideration a large number of variables from multiple sources. Some studies that have used big data methods to define distinct groups of patients (so-called phenomapping) are already available, especially in the fields of oncology, cardiology and diabetes[37,38].

Predictive models. Evidence suggests that early introduction of intensive treatment (that is, combination therapy of biologics and immunosuppressors) in Crohn's disease leads to better outcomes and might be associated with a disease-modifying effect, reducing complications, need for surgery and hospitalizations[39–41]. Features associated with a high risk of an aggressive disease course include perianal disease, ileocolonic location, young age at diagnosis and need for steroids to treat the first flare. However, many patients possess these factors, and they might not be accurate predictors of a severe disease course[42,43]. In ulcerative colitis, factors such as extensive disease, need for systemic corticosteroid therapy at disease onset, young age, extraintestinal manifestations and biochemical parameters were also associated with a more aggressive disease course[44]. Given the potential risks and costs of therapy, defining reliable risk factors and predictive models for severe or complicated disease course in IBD is of paramount importance.

Currently, one the most common uses in health care for big data methodologies is to develop predictive models that identify high-risk or high-cost patients[45], for instance, by including previously unconsidered variables and other difficult-to-handle or complex information such as omics data.

Precision medicine. The IBD therapeutic pipeline has expanded dramatically in the past decade, and several new biologic and small molecule compounds are expected to be available in the next few years[6]. To rationally use these therapeutic resources, it will be crucial to develop biomarkers that reliably identify which patients would benefit, or be harmed, by a particular drug[46]. Efforts have been made to predict response to anti-TNF therapy on the basis of clinical information (such as disease duration, phenotype and smoking status) from retrospective studies and post hoc analyses of clinical trials[47,48], as well as from the study of TNF gene polymorphisms[49,50], but results are inconsistent, and there is a paucity of tools to predict anti-TNF response in clinical practice[51].

In light of this lack of success, tailored therapy for a given patient will probably need input not only from clinical and laboratory information but also from complex omics data. Integration of these multiple data sources in big data studies will therefore be of utmost importance for the development of precision medicine in IBD.

Drug safety. The introduction of new therapies always brings safety concerns, as randomized controlled trials are usually underpowered to detect very infrequent but clinically relevant adverse events. Additionally, such adverse events usually take years or even decades to occur (as in the case of malignancy), beyond the follow-up period of most clinical trials. Currently, the field relies on post-marketing studies, such as the Cancer and Increased Risk Associated with Inflammatory Bowel Disease in France (CESAME)[52] or IBD Cancer and Serious Infections in Europe (I-CARE)[53] studies, but these registries are costly, very time consuming and usually take several years from drug release to develop the full picture of the safety profile of a drug.

By simultaneously evaluating multiple sources of diverse information, big data approaches have the potential to rapidly detect safety signals before currently available tools. Implementation of these techniques applied to drug safety and detection of adverse events is starting to be explored[54–56]. For instance, pharmacovigilance can be improved using text mining, a computational process in which meaningful information is extracted from unstructured textual data sources, to obtain data on adverse drug events from medical notes[54].

Epidemiology and public health. IBD has become a global disease in the past few decades. In developed countries, prevalence is increasing, although the incidence is stable[8]. On the other hand, incidence of IBD in newly industrialized countries has increased steeply, a phenomenon also seen in developing countries with westernization of lifestyle[57]. With this changing epidemiological scenario, the disparity of care across countries will probably be exacerbated[58]. Studies using big data methodologies could help design models that predict health-care utilization to better allocate resources[59,60]. For instance, Sebaa et al. used a Hadoop platform to model equitable health resource allocation in the Béjaïa region in Algeria[59].

Additionally, health-care costs are rapidly increasing worldwide and in the case of IBD are mainly driven by biologic medication costs[61]. In this context, big data research can help improve cost-effectiveness in IBD by correctly identifying patients at risk of an aggressive disease course and those who will benefit from a particular drug at given time of disease.

Drug discovery and development. Although the past decade has seen the IBD pipeline expand markedly, some issues in drug research and development (R&D) still need optimization. The R&D process for new drugs is a very expensive endeavour, ranging from ~US$3 billion to more than $30 billion per approval[62]. Moreover, some compounds prove to be ineffective or even harmful only at late stages of development, wasting great amounts of time and resources and putting individuals at risk. For instance, the antisense oligonucleotide mongersen showed extremely positive effects in a phase II trial in Crohn's disease[63], but the phase III programme was terminated due to futility[64]. In another example, secukinumab, a fully human anti-IL-17A monoclonal antibody, was found to be ineffective, and higher rates of adverse events were noted in the treatment group than in placebo group, despite animal models and genome-wide association studies (GWAS) suggesting a role of IL-17 in Crohn's disease[65]. Tofacitinib, a Janus kinase inhibitor, has also shown inconsistent results in patients with Crohn's disease despite being effective in those with ulcerative colitis[46].

Big data analytics have the potential to improve cost-effectiveness and reduce drug discovery and development times[16,66]. By linking omics data with clinically relevant data from multiple sources, these methods might help prioritize drug targets, mechanisms of action and target populations[67]. Currently, clinical trials need to recruit thousands of patients to develop a drug, and

very frequently, clinical trial results show remarkable variability in responses to a given drug across the studied population. This variability can be explained by omics diversity and phenotypical heterogeneity of the patient population, which can be overturned by the use of big data[68].

Furthermore, big data could be used for repurposing already approved drugs for other indications[16,69,70]. In one example, Dudley et al.[71] applied a computational approach to discover potential new drug therapies for IBD in silico. They compared gene expression profiles from human cell lines treated with 164 different small molecule compounds with publicly available gene expression measurements and data from a previously published study that evaluated Crohn's disease and ulcerative colitis in human intestinal tissue obtained by biopsy[72]. They predicted that the anticonvulsant topiramate would have therapeutic activity in IBD and experimentally validated this finding in vivo in a mouse model[72]. Nevertheless, in a large retrospective cohort study, topiramate use was not associated with a reduction in steroid use, need for anti-TNF agents, surgery or hospitalizations[73], and the drug has not been further investigated in IBD.

Sources of big data in IBD

Administrative databases. Administrative databases are the most straightforward sources to acquire data from for big data research in IBD. Many countries have developed large databases for storing data that are routinely collected during clinic, hospital, laboratory or pharmacy visits[74]. Although most of these databases were initially designed for reimbursement of health-care services, they have been extensively used for epidemiological, effectiveness and safety outcome studies[74].

The French SNIIRAM (Système National d'Information InterRégimes de l'Assurance Maladie) linked with the PMSI (Programme de Médicalisation des Systèmes d'Information) is possibly the world's largest continuous homogeneous claims database[75]. This database includes individual medical and sociodemographic information from all hospital care and outpatient medicine reimbursements of 98.8% of the population living in France (~66 million people) from birth (or immigration) to death (or emigration)[75–78]. The value of this system has been demonstrated in numerous publications, ranging from epidemiological to pharmacoeconomical studies[78], including those in IBD[79,80].

Another European example of a successful administrative database is the British GPRD (General Practice Research Database), a computerized database of anonymized patient data collected continuously since 1987 (REF.[81]). This system contains information on ~4.8 million patients in the United Kingdom, equivalent to ~7% of the population, collected from >600 general practices[81]. The GPRD has proved to be reliable for IBD studies, although it can be difficult to extract relevant information, such as date of incident diagnoses, hospitalizations and surgeries, owing to incomplete records[82].

The Swedish NPR (National Patient Register) was established in 1964 and achieved virtually universal coverage in 2001, when data on specialized hospital-based outpatient care were added[83]. The NPR contains data on diagnoses and procedure codes. The Swedish Quality Register for IBD ($SWIB_{reg}$), established in 2005, contains clinical data that are either missing or lacking in detail in the NPR and covers ~50% of the country's IBD population[84]. Diagnoses of IBD in both the NPR and the $SWIB_{reg}$ have been well validated for use in clinical studies[85]. Notably, many countries across the world have implemented similar databases that enable epidemiological research[86–88].

In the USA, the collection of health data is separated between multiple administrative databases according to specific age or income groups (Medicare and Medicaid services, respectively)[89], profession (for instance, Veterans Affairs)[90] or members of private insurance plans. Often, linkage between different databases or long-term follow-up is not possible. In an effort to homogenize data, a growing number of states have established databases that collect insurance claims information from all health-care payers into all-payer claims databases[91,92], and many other states are considering such a law or programme[91].

Electronic health records. Adoption of electronic health records (EHRs) varies greatly across countries, although rates have been increasing worldwide, and some countries have moved entirely to EHRs[26]. Massive amounts of data are generated and accumulated simply as a by-product of medical attention.

In the USA, physicians have been encouraged to use EHRs since the legislation Health Insurance Portability and Accountability Act was passed in 1996 with the intention to detect insurance fraud[93], but implementation of EHRs varies widely. Adoption of EHRs also varies in Europe, with countries such as Estonia and the Netherlands reaching almost complete coverage[26].

Typically, EHRs include both structured and unstructured data[94]. Structured data account for approximately one-fifth of available information and exist in the form of patient demographics, diagnosis codes, laboratory data, vital signs and similar material. Structured data can be easily stored, analysed and manipulated[18]. However, the vast majority of information in EHRs is unstructured in the form of narrative medical notes[95]; hence, pre-processing of data and computer-based methods such as natural language processing (NLP) are essential to organize, interpret and recognize patterns from these data[94]. In the past 5 years, adoption of NLP in EHR-based research for various purposes, for instance, pharmacovigilance and phenotyping, has grown markedly[96,97]. The performance of NLP has improved greatly and will continue to improve as the number of data sources and their volumes grow[96].

By using data from EHRs, Waljee et al.[98] developed a machine-learning algorithm to predict remission in patients with IBD treated with thiopurines and investigated whether achieving algorithm-predicted remission resulted in fewer clinical events (defined by steroid use, hospitalization or surgery). The algorithm outperformed circulating levels of 6-thioguanine nucleotide in predicting remission (area under the receiver operating characteristic 0.79 versus 0.49), and an algorithm-predicted

Box 1 | **Reasons for big data approaches in IBD research**

- IBDs are heterogeneous diseases that require classification into distinct phenotypes.
- Given the potential risks and cost of therapy, reliable risk predictors and models are needed to implement early disease-modifying strategies.
- Biomarkers will be needed to predict patient response to the growing number of IBD drug classes.
- The safety profile of therapeutic interventions needs to be rapidly defined, especially regarding rare but potentially serious adverse effects.
- IBD epidemiology is rapidly changing, and models are required that predict changes in health-care utilization due to IBD.
- Better strategies are needed to guide drug research, development and repurposing to reduce costs and hasten approvals.

remission was associated with fewer clinical events per year (1.08 versus 3.95; $P < 1 \times 10^{-5}$)[98]. Limitations of this algorithm include the use of retrospective data and a single-centre population in its development, and these results should be validated in prospective trials.

In a study published in 2018, Cai et al. performed a retrospective analysis using NLP to identify arthralgia in the EHR clinical notes from two tertiary hospitals and to compare the risk of arthralgia between patients with IBD receiving vedolizumab and those receiving anti-TNF agents[99]. They found no increased risk of arthralgia associated with vedolizumab use (HR 1.20, 95% CI 0.97–1.49)[99].

Clinical trials and epidemiological studies. Landmark clinical trials have shaped current treatment paradigms in IBD. Moreover, post hoc analyses of these trials have revealed valuable findings, such as the importance of mucosal healing, deep remission and histological remission in disease management. These analyses were mainly reserved for the primary researchers and sponsors; however, there is increasing interest in the need for open-access sharing of data from clinical trials[100]. In 2016, the International Committee of Medical Journal Editors proposed to require authors of clinical trials to share publicly with others the de-identified individual patient data underlying the results presented in the article no later than 6 months after publication to increase the study reproducibility and to facilitate secondary analyses by external investigators[101]. Several factors might hamper the availability of these data, such as intellectual property, fears of different conclusions, confidentiality concerns and lack of resources[102]. Beyond these difficulties, many pharmaceutical sponsors have already created mechanisms for investigators to access patient-level clinical trial data in multiple diseases (including IBD) through open-access platforms[103]. Although the policies by which trials are included in these platforms vary between companies, most include all trials within certain date ranges after regulatory review and publication of results[103]. In an interesting example of how these platforms could enable subsequent analyses, Waljee et al. obtained clinical data from the induction and maintenance phase III trial of vedolizumab in ulcerative colitis (GEMINI 1) via the Clinical Study Data Request open-access platform[104]. They then applied machine-learning tools to develop predictive models of corticosteroid-free endoscopic remission in response to vedolizumab[105]. Although open data platforms are an opportunity for research, with data available from >3,000 trials, they are underutilized: only 15.7% of trial data sets had been requested by a limited number of researchers as of 2016 (REF.[103]).

Epidemiological studies such as the IBSEN study and the CESAME study have also greatly contributed to the understanding of IBD, especially regarding natural history and safety of interventions[52,106–108]. Examples of future cohort studies include the I-CARE (NCT02377258, which will look deeper into the risk of malignancy and infections)[53,109] and the PREdiCCt studies (NCT03282903)[110]. For instance, in the PREdiCCT study, patient-generated data on clinical symptoms, diet and lifestyle gathered through a mobile application[110] will be integrated with genomic and microbiota data in a multisource input paradigm to study the effects of these factors on IBD flares and recovery[110].

The main strength of the information gathered in clinical trials and cohort studies for big data analytics is its high quality and consistency, whereas the availability of data represents the main limitation. In turn, big data approaches might help the design of both interventional and observational clinical studies, such as by improving trial designs, tailoring patient selection, boosting recruitment and lowering costs[111,112].

Mobile applications, e-health and social media. During the past two decades, a remarkable shift has occurred towards the digitalization of daily life. The internet and mobile technologies are present in almost every aspect of life, with social media having a preponderant role[113]. The 'read-only' World Wide Web environment has evolved to Web 2.0, characterized by multidirectional communication in which individuals produce, participate, modify and collaborate with user-generated content[114,115]. These digital interactions lead to the accumulation of an enormous amount of data. e-Health tools and telemedicine (defined as diagnosis, treatment and monitoring of disease at a distance, especially by means of the internet, mobile phone applications and wearable devices) not only arise as a consequence of this context but might also be an opportunity to facilitate self-management and reduce health-care utilization[116,117].

The effect of e-health in IBD has been studied in a few clinical trials with dissimilar results[118–121]. Whereas earlier trials showed the value of these strategies only in patients with ulcerative colitis (mainly those with mild to moderate disease)[122,123], a large randomized controlled trial conducted in Netherlands and published in 2017 demonstrated that a telemedicine system through a web-based and smartphone application was efficacious in all subtypes of IBD[119]. Those in the intervention group had reduced use of health-care services (number of outpatient visits and hospital admissions) and increased treatment adherence compared with patients in the standard care group[119]. However, another randomized controlled trial published in 2018 showed no differences in disease activity and quality of life between telemedicine and standard of care groups after 1 year; telemedicine was associated with a decrease in hospitalizations but also with an

overall increase in health-care utilization[124]. A comprehensive telemedicine system in IBD should include not only patient-reported outcome data but also objective markers of inflammation[125]. In this regard, faecal calprotectin levels measured using a home-based test linked to a smartphone application showed good correlation with levels determined by laboratory-based enzyme-linked immunosorbent assay (ELISA) analysis[126,127]. The implementation of e-health, and its use as a source of big data analytics will surely face challenges, especially regarding data privacy, security and legal ownership[125].

Social media can also serve as a data source that offers particular opportunities to gain new insight on health-seeking behaviour, epidemiological trends and patients' perspectives of disease and treatments[128]. For instance, a study published in 2017 used a netnography analysis — a method to understand social interactions in the context of contemporary social networks — to evaluate posts from Twitter and >3,000 social media sites to reveal patients' experience and choice of biologics in IBD[129]. They examined 1,598 IBD-related posts and found that the main themes of interaction were negative experiences with biologics, decision-making surrounding biologic use, positive experiences with biologics, information-seeking from peers and costs[129].

Medical imaging. Imaging techniques, particularly MRI, CT and ultrasonography, are increasingly used as diagnostic tools and non-invasive objective measures of inflammation in IBD[130]. As these techniques become widely available and cloud systems are used to digitally store and process these imaging study findings, the volume of data in the form of medical images will continue to grow exponentially[131]. Application of big data methodologies in the field of medical imaging has the potential to enhance pattern recognition of lesions to have more accurate interpretation of results. Big data can also help determine which patients will have a better diagnostic yield for a given imaging technique[132,133]. Challenges of its use include the difficulty of comparing images obtained using different techniques and integration of imaging data with other sources.

Genomics, proteomics, metabolomics and microbiomics. GWAS have identified multiple loci associated with increased risk of IBD[134–136]. Moreover, high-resolution genetic studies have identified within these loci the specific single nucleotide variants (SNVs) responsible for the increase in IBD risk[137], although the underlying mechanisms linking individual SNVs to disease risk are still unclear. Some genetic variants proved to be associated with distinct disease phenotypes, such as *NOD2* gene mutations in fibrostenotic Crohn's disease[138]. However, as a general rule, most genetic variants have a rather small effect on overall disease risk, prognosis or response to therapy, implying that genetic variants are by themselves not predictive and that most people carrying a high-risk variant will never develop the disease[139,140]. Moreover, most of these risk variants are shared with other chronic inflammatory diseases, such as mutations in *IL23R* in ankylosing spondylitis and psoriasis, and mutations in *NOD2* in mycobacterial disease[141], which indicates that, although they might contribute to an overall increase in inflammatory disease risk, they do not dictate organ specificity[142,143]. Overall, these data imply that the phenotypic effect of genetic variants is modulated by a plethora of non-genetic factors, which probably include the diet as well as the composition and diversity of the intestinal microbiome[144]. The role of these additional factors imposes the need to integrate data from GWAS with data from other omics approaches, such as those investigating changes in gene expression and the accessibility and usage of the genome (for example, changes in DNA methylation) in both intestinal and immune cells. Efforts in this direction are now being carried out worldwide in large-scale consortia projects, such as the Systems Medicine Approach to Chronic Inflammatory Diseases (SYSCID) consortium[145].

Owing to the rapid generation of enormous amounts of omics data in the past decade, problems related to storage, analysis, integration and interpretation have arisen[146,147] that have largely been solved by computational techniques using algorithmic frameworks that are adaptable to large-scale omics data[148]. It is now clear that bioinformatics and computational sciences are essential to adequately manage and integrate data from these components and other sources[3,149,150].

Table 1 | **Examples of big data studies in IBD**

Study	Big data method	Application
Waljee A. K. et al.[98], Waljee A. K. et al.[153]	Machine-learning model using EHR data	Identification of objective remission in patients with IBD treated with thiopurines
Waljee A. K. et al.[154]	Machine-learning model using EHR data	Prediction of outpatient corticosteroid use and hospitalization
Wei Z. et al.[155]	Machine-learning model using data set from the International IBD Genetics Consortium	Risk prediction for UC and CD
Waljee A. K. et al.[105]	Machine-learning model using data from GEMINI 1 clinical trial	Prediction of corticosteroid-free remission with vedolizumab in patients with UC
Menti E. et al.[156]	Bayesian machine-learning model using clinical, phenotypical and genetic data	Risk prediction for extraintestinal manifestations in CD
Han L. et al.[157]	Gaussian Bayesian network	Differentiation between CD and UC
Cai T. et al.[99]	Natural language processing of EHR data	Identification of arthralgia in patients with IBD treated with vedolizumab
Hou J. K. et al.[158]	Natural language processing of EHR data	Differentiation of surveillance and non-surveillance colonoscopies in patients with IBD

CD, Crohn's disease; EHR: electronic health record; UC, ulcerative colitis.

Conclusions

Most aspects of life have become increasingly digitized over the past few years. Data are generated and accumulated simply as a by-product, and the health-care sector is no exception to this fact. Enormous amounts of data are

Box 2 | **Potential limitations and challenges of big data research in IBD**

Data heterogeneity
For example, social media posts and unstructured electronic health record notes

Poor quality data
For example, corrupted, duplicate, missing or inaccurate data

Ethical and legal constraints to data availability
Some data sources raise issues of patient privacy and/or consent, data security, intellectual property and protection of commercial interests, among others

Need for clinical validation of prediction models
Risk models still need to be validated in clinical trials

Integration in clinical practice
Models should prove their worth in real-world settings

generated through various sources, such as EHRs, administrative databases, clinical trials, registries, social media and omics techniques. Big data studies are an important opportunity to leverage these underutilized data sources and to gain new insights that ultimately lead to better understanding of IBD and fill the gaps in patient care.

IBD research has seen great advances, although clearly, there are many unmet needs (BOX 1). Currently, the most potent biologic treatments benefit roughly half of patients at most, and complications, impaired quality of life, hospitalizations and surgeries are still common. Despite the introduction of biosimilar agents, treatment-related costs are still very high, and in the context of increasing incidence in low-income and middle-income countries, improved IBD care cost-effectiveness is an important goal.

Implementation of big data methodologies in IBD research is very promising (TABLE 1), but it must be remembered that these research strategies are at early stages in health care in general. Even in pioneer disciplines in the field, such as oncology and cardiology, the reports are scant, and the added value of big data remains to be seen. Lack of direct evidence and the disappointing results of initial studies (such as the aforementioned Google influenza model) urge caution.

Researchers will face several limitations and challenges with the implementation of big data approaches in IBD (BOX 2). First, the quality of data across different sources will inherently be heterogeneous, with some sources (for instance, social media or even unstructured information in EHRs) especially prone to poor quality[151]. Big data approaches can be performed with poor data quality inputs, which can detrimentally affect the accuracy and clinical utility of the output[18]. Identification and selection of correct and adequate-quality sources represent important challenges to achieve a critical characteristic of big data in health care: veracity. Second, the availability of data faces ethical and legal constraints related to patient privacy and consent to share individual data. Although, personal information is de-identified when data are analysed, the possibility of recognizing individuals still exists[152]. Third, predictions and models made by computational methods must still be thoroughly validated experimentally and clinically before general use[16], as poorly validated models might have the potential to harm[151]. Independent agencies must oversee and certify commercial profit-driven initiatives that intend to be used in clinical practice[151]. Fourth, to potentially improve disease management and outcomes, big data outputs must be integrated into clinical practice, and the question of whether big data models are more effective than traditional risk models remains to be seen[17].

Big data research has overcome some of these challenges and proved its value in other fields, such as finance and politics. The era of big data in health care is definitely still in its infancy, but hopefully, IBD research will benefit from its many promises in the coming years.

Published online 18 January 2019

1. Ungaro, R., Mehandru, S., Allen, P. B., Peyrin-Biroulet, L. & Colombel, J.-F. Ulcerative colitis. *Lancet* **389**, 1756–1770 (2017).
2. Torres, J., Mehandru, S., Colombel, J.-F. & Peyrin-Biroulet, L. Crohn's disease. *Lancet* **389**, 1741–1755 (2016).
3. de Souza, H. S. P., Fiocchi, C. & Iliopoulos, D. The IBD interactome: an integrated view of aetiology, pathogenesis and therapy. *Nat. Rev. Gastroenterol. Hepatol.* **14**, 739–749 (2017).
4. Jin, L. et al. Pathway-based analysis tools for complex diseases: a review. *Genomics Proteomics Bioinformatics* **12**, 210–220 (2014).
5. Actis, G. C. & Rosina, F. Inflammatory bowel disease: an archetype disorder of outer environment sensor systems. *World J. Gastrointest. Pharmacol. Ther.* **4**, 41–46 (2013).
6. Olivera, P., Danese, S. & Peyrin-Biroulet, L. Next generation of small molecules in inflammatory bowel disease. *Gut* **66**, 199–209 (2017).
7. Ng, S. C. et al. Environmental risk factors in inflammatory bowel disease: a population-based case-control study in Asia-Pacific. *Gut* **64**, 1063–1071 (2015).
8. Ng, S. C. et al. Worldwide incidence and prevalence of inflammatory bowel disease in the 21st century: a systematic review of population-based studies. *Lancet* **390**, 2769–2778 (2017).
9. Bernstein, C. N. Treatment of IBD: where we are and where we are going. *Am. J. Gastroenterol.* **110**, 114–126 (2015).
10. Actis, G. C., Pellicano, R. & Rosina, F. Inflammatory bowel diseases: current problems and future tasks. *World J. Gastrointest. Pharmacol. Ther.* **5**, 169–174 (2014).
11. Manyika, J. et al. *Big Data: The Next Frontier for Innovation, Competition, and Productivity* (McKinsey, 2011).
12. Philip Chen, C. L. & Zhang, C.-Y. Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. *Inf. Sci.* **275**, 314–347 (2014).
13. Nickerson, D. W. & Rogers, T. Political campaigns and big data. *J. Econ. Perspect.* **28**, 51–74 (2014).
14. Kayyali, B., Knott, D. & Van Kuiken, S. The big-data revolution in US health care: accelerating value and innovation. *McKinsey* https://www.mckinsey.com/industries/healthcare-systems-and-services/our-insights/the-big-data-revolution-in-us-health-care (2013).
15. Obermeyer, Z. & Emanuel, E. J. Predicting the future — big data, machine learning, and clinical medicine. *N. Engl. J. Med.* **375**, 1216–1219 (2016).
16. Wooden, B., Goossens, N., Hoshida, Y. & Friedman, S. L. Using big data to discover diagnostics and therapeutics for gastrointestinal and liver diseases. *Gastroenterology* **152**, 53–67 (2017).
17. Rumsfeld, J. S., Joynt, K. E. & Maddox, T. M. Big data analytics to improve cardiovascular care: promise and challenges. *Nat. Rev. Cardiol.* **13**, 350–359 (2016).
18. Raghupathi, W. & Raghupathi, V. Big data analytics in healthcare: promise and potential. *Health Inf. Sci. Syst.* **2**, 3 (2014).
19. Alonso, S. G., de la Torre Díez, I., Rodrigues, J. J. P. C., Hamrioui, S. & López-Coronado, M. A. systematic review of techniques and sources of big data in the healthcare sector. *J. Med. Syst.* **41**, 183 (2017).
20. Ginsberg, J. et al. Detecting influenza epidemics using search engine query data. *Nature* **457**, 1012–1014 (2009).
21. Butler, D. When Google got flu wrong. *Nature* **494**, 155–156 (2013).
22. Lazer, D., Kennedy, R., King, G. & Vespignani, A. The parable of Google Flu: traps in big data analysis. *Science* **343**, 1203–1205 (2014).
23. Dinov, I. D. et al. Predictive big data analytics: a study of Parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PLOS ONE* **11**, 1–28 (2016).
24. US National Institutes of Health. Big data for knowledge. *NIH* https://commonfund.nih.gov/bd2k (2018).
25. Ketchersid, T. Big data in nephrology: friend or foe? *Blood Purif.* **36**, 160–164 (2014).
26. Auffray, C. et al. Making sense of big data in health research: towards an EU action plan. *Genome Med.* **8**, 71 (2016).
27. Bellazzi, R. Big data and biomedical informatics: a challenging opportunity. *Yearb. Med. Inform.* **9**, 8–13 (2014).
28. Corbin, K. How CIOs can prepare for healthcare 'data tsunami'. *CIO* https://www.cio.com/article/2860072/healthcare/how-cios-can-prepare-for-healthcare-data-tsunami.html (2014).
29. Lee, C. H. & Yoon, H.-J. Medical big data: promise and challenges. *Kidney Res. Clin. Pract.* **36**, 3–11 (2017).
30. Weber, G. M., Mandl, K. D. & Kohane, I. S. Finding the missing link for big biomedical data. *JAMA* **311**, 2479–2480 (2014).
31. Carroll, S. & Goodstein, D. Defining the scientific method. *Nat. Methods* **6**, 237–237 (2009).
32. Subramanian, S., Ekbom, A. & Rhodes, J. M. Recent advances in clinical practice: a systematic review of isolated colonic Crohn's disease: the third IBD? *Gut* **66**, 362–381 (2017).

33. Ruel, J., Ruane, D., Mehandru, S., Gower-Rousseau, C. & Colombel, J.-F. IBD across the age spectrum: is it the same disease? *Nat. Rev. Gastroenterol. Hepatol.* **11**, 88–98 (2014).
34. Aloi, M. et al. Phenotype and disease course of early-onset pediatric inflammatory bowel disease. *Inflamm. Bowel Dis.* **20**, 597–605 (2014).
35. Billiet, T. & Vermeire, S. Differences between adults and children: genetics and beyond. *Expert Rev. Gastroenterol. Hepatol.* **9**, 191–196 (2015).
36. Peyrin-Biroulet, L. et al. Defining disease severity in inflammatory bowel diseases: current and future directions. *Clin. Gastroenterol. Hepatol.* **14**, 348–354 (2016).
37. Shivade, C. et al. A review of approaches to identifying patient phenotype cohorts using electronic health records. *J. Am. Med. Inform. Assoc.* **21**, 221–230 (2014).
38. Altman, R. B. & Ashley, E. A. Using "big data" to dissect clinical heterogeneity. *Circulation* **131**, 232–233 (2015).
39. D'Haens, G. et al. Early combined immunosuppression or conventional management in patients with newly diagnosed Crohn's disease: an open randomised trial. *Lancet* **371**, 660–667 (2008).
40. Peyrin-Biroulet, L. et al. Impact of azathioprine and tumour necrosis factor antagonists on the need for surgery in newly diagnosed Crohn's disease. *Gut* **60**, 930–936 (2011).
41. Allen, P. B. et al. Review article: moving towards common therapeutic goals in Crohn's disease and rheumatoid arthritis. *Aliment. Pharmacol. Ther.* **45**, 1058–1072 (2017).
42. Gomollón, F. et al. 3rd European evidence-based consensus on the diagnosis and management of Crohn's disease 2016 —part 1: diagnosis and medical management. *J. Crohns Colitis* **11**, 3–25 (2017).
43. Cosnes, J. et al. Early administration of azathioprine versus conventional management of Crohn's Disease: a randomized controlled trial. *Gastroenterology* **145**, 758–765 (2013).
44. Stallmach, A. et al. Parameters of a severe disease course in ulcerative colitis. *World J. Gastroenterol.* **20**, 12574–12580 (2014).
45. Bates, D. W., Saria, S., Ohno-Machado, L., Shah, A. & Escobar, G. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff.* **33**, 1123–1131 (2014).
46. Olivera, P., Danese, S. & Peyrin-Biroulet, L. JAK inhibition in inflammatory bowel disease. *Expert Rev. Clin. Immunol.* **13**, 693–703 (2017).
47. Chaudhary, R. & Ghosh, S. Prediction of response to infliximab in Crohn's disease. *Dig. Liver Dis.* **37**, 559–563 (2005).
48. Siegel, C. A. & Melmed, G. Y. Predicting response to Anti-TNF Agents for the treatment of crohn's disease. *Therap. Adv. Gastroenterol.* **2**, 245–251 (2009).
49. Arijs, I. et al. Mucosal gene signatures to predict response to infliximab in patients with ulcerative colitis. *Gut* **58**, 1612–1619 (2009).
50. Burke, K. E. et al. Genetic markers predict primary nonresponse and durable response to anti-tumor necrosis factor therapy in ulcerative colitis. *Inflamm. Bowel Dis.* **24**, 1840–1848 (2018).
51. Boyapati, R. K., Kalla, R., Satsangi, J. & Ho, G. Biomarkers in search of precision medicine in IBD. *Am. J. Gastroenterol.* **111**, 1682–1690 (2016).
52. Beaugerie, L. et al. Lymphoproliferative disorders in patients receiving thiopurines for inflammatory bowel disease: a prospective observational cohort study. *Lancet* **374**, 1617–1625 (2009).
53. The I-CARE Study Group. P509 IBD cancer and serious infections in Europe (I-CARE): a European prospective observational study. *J. Crohns Colitis* **11**, S338–S339 (2017).
54. Harpaz, R. et al. Text mining for adverse drug events: the promise, challenges, and state of the art. *Drug Saf.* **37**, 777–790 (2014).
55. Arnaud, M. et al. Methods for safety signal detection in healthcare databases: a literature review. *Expert Opin. Drug Saf.* **16**, 721–732 (2017).
56. Wang, G., Jung, K., Winnenburg, R. & Shah, N. H. A method for systematic discovery of adverse drug events from clinical notes. *J. Am. Med. Inform. Assoc.* **22**, 1196–1204 (2015).
57. Kaplan, G. G. & Ng, S. C. Understanding and preventing the global increase of inflammatory bowel disease. *Gastroenterology* **152**, 313–321 (2017).
58. Kaplan, G. G. The global burden of IBD: from 2015 to 2025. *Nat. Rev. Gastroenterol. Hepatol.* **12**, 720–727 (2015).
59. Sebaa, A., Chikh, F., Nouicer, A. & Tari, A. Medical big data warehouse: architecture and system design, a case study: improving healthcare resources distribution. *J. Med. Syst.* **42**, 59 (2018).
60. Bram, J. T., Warwick-Clark, B., Obeysekare, E. & Mehta, K. Utilization and monetization of healthcare data in developing countries. *Big Data* **3**, 59–66 (2015).
61. van der Valk, M. E. et al. Healthcare costs of inflammatory bowel disease have shifted from hospitalisation and surgery towards anti-TNFα therapy: results from the COIN study. *Gut* **63**, 72–79 (2014).
62. Schuhmacher, A., Gassmann, O. & Hinder, M. Changing R&D models in research-based pharmaceutical companies. *J. Transl Med.* **14**, 105 (2016).
63. Monteleone, G. et al. Mongersen, an oral SMAD7 antisense oligonucleotide, and Crohn's disease. *N. Engl. J. Med.* **372**, 1104–1113 (2015).
64. Celgene. Celgene provides update on GED-0301 (mongersen) inflammatory bowel disease program. *Celgene* https://ir.celgene.com/press-releases/press-release-details/2017/Celgene-Provides-Update-on-GED-0301-mongersen-Inflammatory-Bowel-Disease-Program/default.aspx (2017).
65. Hueber, W. et al. Secukinumab, a human anti-IL-17A monoclonal antibody, for moderate to severe Crohn's disease: unexpected results of a randomised, double-blind placebo-controlled trial. *Gut* **61**, 1693–1700 (2012).
66. Chen, B. & Butte, A. J. Leveraging big data to transform target selection and drug discovery. *Clin. Pharmacol. Ther.* **99**, 285–297 (2016).
67. Denny, J. C., Van Driest, S. L., Wei, W. Q. & Roden, D. M. The influence of big (clinical) data and genomics on precision medicine and drug development. *Clin. Pharmacol. Ther.* **103**, 409–418 (2018).
68. Blackburn, M., Alexander, J., Legan, J. D. & Klabjan, D. Big data and the future of R&D management: the rise of big data and big data analytics will have significant implications for R&D and innovation management in the next decade. *Res. Technol. Manag.* **60**, 43–51 (2017).
69. Power, A., Berger, A. C. & Ginsburg, G. S. Genomics-enabled drug repositioning and repurposing. *JAMA* **311**, 2063 (2014).
70. Li, J. et al. A survey of current trends in computational drug repositioning. *Brief. Bioinform.* **17**, 2–12 (2016).
71. Lamb, J. et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science* **313**, 1929–1935 (2006).
72. Dudley, J. T. et al. Computational repositioning of the anticonvulsant topiramate for inflammatory bowel disease. *Sci. Transl Med.* **3**, 96ra76 (2011).
73. Crockett, S. D., Schectman, R., Stürmer, T. & Kappelman, M. D. Topiramate use does not reduce flares of inflammatory bowel disease. *Dig. Dis. Sci.* **59**, 1535–1543 (2014).
74. Hashimoto, R. E., Brodt, E. D., Skelly, A. C. & Dettori, J. R. Administrative database studies: goldmine or goose chase? *Evid. Based Spine Care J.* **5**, 74–76 (2014).
75. Bezin, J. et al. The national healthcare system claims databases in France, SNIIRAM and EGB: powerful tools for pharmacoepidemiology. *Pharmacoepidemiol. Drug Saf.* **26**, 954–962 (2017).
76. Moulis, G. et al. French health insurance databases: what interest for medical research? *Rev. Med. Interne* **36**, 411–417 (2015).
77. Tuppin, P., de Roquefeuil, L., Weill, A., Ricordeau, P. & Merlière, Y. French national health insurance information system and the permanent beneficiaries sample. *Rev. Epidemiol. Sante Publique* **58**, 286–290 (2010).
78. Tuppin, P. et al. Value of a national administrative database to guide public decisions: from the système national d'information interrégimes de l'Assurance Maladie (SNIIRAM) to the système national des données de santé (SNDS) in France. *Rev. Epidemiol. Sante Publique* **65**, S149–S167 (2017).
79. Blotière, P. O. et al. Conditions of prescription of anti-TNF agents in newly treated patients with inflammatory bowel disease in France (2011–2013). *Dig. Liver Dis.* **48**, 620–625 (2016).
80. Lemaitre, M. et al. Association between use of thiopurines or tumor necrosis factor antagonists alone or in combination and risk of lymphoma in patients with inflammatory bowel disease. *JAMA* **318**, 1679 (2017).
81. Medicines and Healthcare products Regulatory Agency. The General Practice Research Database (GPRD) — further infromation for patients. *NHS Scotland* http://www.erskinepractice.scot.nhs.uk/website/S11486/files/GPRD_PatientLeaflet.pdf (2010).
82. Lewis, J. D., Brensinger, C., Bilker, W. B. & Strom, B. L. Validity and completeness of the General Practice Research Database for studies of inflammatory bowel disease. *Pharmacoepidemiol. Drug Saf.* **11**, 211–218 (2002).
83. Ludvigsson, J. F. et al. External review and validation of the Swedish national inpatient register. *BMC Public Health* **11**, 450 (2011).
84. SWIBREG. Swedish Inflammatory Bowel Disease Registry. *SWIBREG* http://www.swibreg.se/ (2018).
85. Jakobsson, G. L. et al. Validating inflammatory bowel disease (IBD) in the Swedish National Patient Register and the Swedish Quality Register for IBD (SWIBREG). *Scand. J. Gastroenterol.* **52**, 216–221 (2017).
86. Schmidt, M. et al. The Danish National patient registry: a review of content, data quality, and research potential. *Clin. Epidemiol.* **7**, 449–490 (2015).
87. Kreis, K., Neubauer, S., Klora, M., Lange, A. & Zeidler, J. Status and perspectives of claims data analyses in Germany — a systematic review. *Health Policy* **120**, 213–226 (2016).
88. Cheng, C.-L. et al. Validation of acute myocardial infarction cases in the national health insurance research database in taiwan. *J. Epidemiol.* **24**, 500–507 (2014).
89. Lichtman, J. H., Leifheit-Limson, E. C. & Goldstein, L. B. Centers for medicare and medicaid services medicare data and stroke research: goldmine or landmine? *Stroke* **46**, 598–604 (2015).
90. Boyko, E. J., Koepsell, T. D., Gaziano, J. M., Horner, R. D. & Feussner, J. R. US Department of Veterans Affairs medical care system as a resource to epidemiologists. *Am. J. Epidemiol.* **151**, 307–314 (2000).
91. National Conference of State Legislatures. Collecting health data: all-payers claims databases. *NCSL* http://www.ncsl.org/research/health/collecting-health-data-all-payer-claims-database.aspx (2018).
92. All-Payer Claims Database Council. APCD Council. *APCD Council* https://www.apcdcouncil.org/ (2018).
93. US Department of Health & Human Services. Health information privacy. *HHS* https://www.hhs.gov/hipaa/index.html (2018).
94. Ross, M. K., Wei, W. & Ohno-Machado, L. "Big data" and the electronic health record. *Yearb. Med. Inform.* **9**, 97–104 (2014).
95. Austin, C. & Kusumoto, F. The application of Big Data in medicine: current implications and future directions. *J. Interv. Card. Electrophysiol.* **47**, 51–59 (2016).
96. Luo, Y. et al. Natural language processing for EHR-based pharmacovigilance: a structured review. *Drug Saf.* **40**, 1075–1089 (2017).
97. Zeng, Z., Deng, Y., Li, X., Naumann, T. & Luo, Y. Natural language processing for EHR-based computational phenotyping. *IEEE ACM Trans. Comput. Biol. Bioinform.* https://doi.org/10.1109/TCBB.2018.2849968 (2018).
98. Waljee, A. K. et al. Machine learning algorithms for objective remission and clinical outcomes with thiopurines. *J. Crohns Colitis* **108**, 1723–1730 (2017).
99. Cai, T. et al. The association between arthralgia and vedolizumab using natural language processing. *Inflamm. Bowel Dis.* **24**, 2242–2246 (2018).
100. Krumholz, H. M. & Peterson, E. D. Open access to clinical trials data. *JAMA* **312**, 1002–1003 (2014).
101. Taichman, D. B. et al. Sharing clinical trial data — a proposal from the International Committee of Medical Journal Editors. *N. Engl. J. Med.* **374**, 384–386 (2016).
102. Bertagnolli, M. M. et al. Advantages of a truly open-access data-sharing model. *N. Engl. J. Med.* **376**, 1178–1181 (2017).
103. Navar, A. M., Pencina, M. J., Rymer, J. A., Louzao, D. M. & Peterson, E. D. Use of open access platforms for clinical trial data. *JAMA* **315**, 1283 (2016).
104. Clinical Study Data Request. Clinical Study Data Request. *CSDR* https://clinicalstudydatarequest.com/Default.aspx (2018).
105. Waljee, A. K. et al. Predicting corticosteroid-free endoscopic remission with vedolizumab in ulcerative colitis. *Aliment. Pharmacol. Ther.* **47**, 763–772 (2018).
106. Elriz, K. et al. Incidence, presentation, and prognosis of small bowel adenocarcinoma in patients with small bowel Crohn's disease: a prospective observational study. *Inflamm. Bowel Dis.* **19**, 1823–1826 (2013).

107. Henriksen, M. et al. Ulcerative colitis and clinical course: results of a 5-year population-based follow-up study (the IBSEN study). *Inflamm. Bowel Dis.* **12**, 543–550 (2006).
108. Hovde, Ø. et al. Malignancies in patients with inflammatory bowel disease: results from 20 years of follow-up in the IBSEN study. *J. Crohns Colitis* **11**, 571–577 (2016).
109. US National Library of Medicine. *ClinicalTrials.gov* https://www.clinicaltrials.gov/ct2/show/NCT02377258 (2016).
110. US National Library of Medicine. *ClinicalTrials.gov* https://www.clinicaltrials.gov/ct2/show/NCT03282903 (2018).
111. Mayo, C. S. et al. Big data in designing clinical trials: opportunities and challenges. *Front. Oncol.* **7**, 187 (2017).
112. Angus, D. C. Fusing randomized trials with big data: the key to self-learning health care systems? *JAMA* **314**, 767–768 (2015).
113. Perrin, A. Social media usage: 2005–2015. *Pew Research Center* http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/ (2015).
114. Veríssimo, J. M. C. Usage intensity of mobile medical apps: a tale of two methods. *J. Bus. Res.* **89**, 442–447 (2017).
115. Chou, W. S., Prestin, A., Lyons, C. & Wen, K. Web 2.0 for health promotion: reviewing the current evidence. *Am. J. Public Health* **103**, e9–e18 (2013).
116. Orozco-Beltran, D., Sánchez-Molla, M., Sanchez, J. J. & Mira, J. J., ValCrònic Research Group. Telemedicine in primary care for patients with chronic conditions: the ValCrònic Quasi-Experimental Study. *J. Med. Internet Res.* **19**, e400 (2017).
117. Jackson, B. D., Con, D. & De Cruz, P. Design considerations for an eHealth decision support tool in inflammatory bowel disease self-management. *Intern. Med. J.* **48**, 674–681 (2017).
118. Boelle, P.-Y., Thiébaut, R. & Costagliola, D. Données massives, vous avez dit données massives? *Quest. Santé Publique* **30**, 1–4 (2015).
119. De Jong, M. et al. Development and feasibility study of a telemedicine tool for all patients with IBD: MyIBDcoach. *Inflamm. Bowel Dis.* **23**, 485–493 (2017).
120. Jackson, B. D., Gray, K., Knowles, S. R. & De Cruz, P. EHealth technologies in inflammatory bowel disease: a systematic review. *J. Crohns Colitis* **10**, 1103–1121 (2016).
121. Jaboli, F., Pouillon, L., Bossuyt, P., Danese, S. & Peyrin-Biroulet, L. Telehealth in inflammatory bowel disease: every patient may need a coach! *Gastroenterology* **154**, 1196–1198 (2018).
122. Cross, R. K., Cheevers, N., Rustgi, A., Langenberg, P. & Finkelstein, J. Randomized, controlled trial of home telemanagement in patients with ulcerative colitis (UC HAT). *Inflamm. Bowel Dis.* **18**, 1018–1025 (2012).
123. Elkjaer, M. et al. E-Health empowers patients with ulcerative colitis: a randomised controlled trial of the web-guided "Constant-care" approach. *Gut* **59**, 1652–1661 (2010).
124. Cross, R. K. et al. A randomized controlled trial of telemedicine for patients with inflammatory bowel disease (Tele-IBD) [abstract 903]. *Gastroenterology* **154**, S177 (2018).
125. Bossuyt, P., Pouillon, L. & Peyrin-Biroulet, L. Primetime for e-health in IBD? *Nat. Rev. Gastroenterol. Hepatol.* **14**, 133–134 (2017).
126. Bello, C. et al. Usability of a home-based test for the measurement of fecal calprotectin in asymptomatic IBD patients. *Dig. Liver Dis.* **49**, 991–996 (2017).
127. Heida, A. et al. Agreement between home-based measurement of stool calprotectin and ELISA results for monitoring inflammatory bowel disease activity. *Clin. Gastroenterol. Hepatol.* **15**, 1742–1749 (2017).
128. Mgudlwa, S. & Iyamu, T. Integration of social media with healthcare big data for improved service delivery. *SA J. Inf. Manag.* **20**, 1–8 (2018).
129. Martinez, B. et al. Patient understanding of the risks and benefits of biologic therapies in inflammatory bowel disease: insights from a large-scale analysis of social media platforms. *Inflamm. Bowel Dis.* **23**, 1057–1064 (2017).
130. Panes, J. et al. Imaging techniques for assessment of inflammatory bowel disease: joint ECCO and ESGAR evidence-based consensus guidelines. *J. Crohns Colitis* **7**, 556–585 (2013).
131. Belle, A. et al. Big data analytics in healthcare. *Biomed. Res. Int.* **2015**, 370194 (2015).
132. Dilsizian, S. E. & Siegel, E. L. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Curr. Cardiol. Rep.* **16**, 441 (2014).
133. Landewé, R. B. M. & van der Heijde, D. "Big data" in rheumatology: intelligent data modeling improves the quality of imaging data. *Rheum. Dis. Clin. North Am.* **44**, 307–315 (2018).
134. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
135. Anderson, C. A. et al. Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. *Nat. Genet.* **43**, 246–252 (2011).
136. Franke, A. et al. Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. *Nat. Genet.* **42**, 1118–1125 (2010).
137. Huang, H. et al. Fine-mapping inflammatory bowel disease loci to single-variant resolution. *Nature* **547**, 173–178 (2017).
138. Abreu, M. T. et al. Mutations in NOD2 are associated with fibrostenosing disease in patients with Crohn's disease. *Gastroenterology* **123**, 679–688 (2002).
139. Ellinghaus, D., Bethune, J., Petersen, B. S. & Franke, A. The genetics of Crohn's disease and ulcerative colitis-status quo and beyond. *Scand. J. Gastroenterol.* **50**, 13–23 (2014).
140. Mirkov, M. U., Verstockt, B. & Cleynen, I. Genetics of inflammatory bowel disease: beyond NOD2. *Lancet Gastroenterol. Hepatol.* **2**, 224–234 (2017).
141. Lees, C. W., Barrett, J. C., Parkes, M. & Satsangi, J. New IBD genetics: common pathways with other diseases. *Gut* **60**, 1739–1753 (2011).
142. Ye, B. D. & McGovern, D. P. B. Genetic variation in IBD: progress, clues to pathogenesis and possible clinical utility. *Expert Rev. Clin. Immunol.* **12**, 1091–1107 (2016).
143. Ellinghaus, D. et al. Analysis of five chronic inflammatory diseases identifies 27 new associations and highlights disease-specific patterns at shared loci. *Nat. Genet.* **48**, 510–518 (2016).
144. Zuo, T., Kamm, M. A., Colombel, J. F. & Ng, S. C. Urbanization and the gut microbiota in health and inflammatory bowel disease. *Nat. Rev. Gastroenterol. Hepatol.* **15**, 440–452 (2018).
145. SYSCID. SYSCID — a systems medicine approach to chronic inflammatory diseases. *SYSCID* https://syscid.eu (2018).
146. Schultze, J. L. & Rosenstiel, P. The SYSCID Consortium. Systems medicine in chronic inflammatory diseases. *Immunity* **48**, 608–613 (2018).
147. Gedela, S. Integration, warehousing, and analysis strategies of Omics data. *Methods Mol. Biol.* **719**, 399–414 (2011).
148. Berger, B., Peng, J. & Singh, M. Computational solutions for omics data. *Nat. Rev. Genet.* **14**, 333–346 (2013).
149. Fiocchi, C. Integrating omics: the future of IBD? *Dig. Dis.* **32**, 96–102 (2014).
150. Chuong, K. H., Mack, D. R., Stintzi, A. & O'Doherty, K. C. Human microbiome and learning healthcare systems: integrating research and precision medicine for inflammatory bowel disease. *OMICS* **22**, 119–126 (2017).
151. Shah, N. D. et al. Big data and predictive analytics recalibrating expectations. *JAMA* **320**, 27–28 (2018).
152. Genta, R. M. & Sonnenberg, A. Big data in gastroenterology research. *Nat. Rev. Gastroenterol. Hepatol.* **11**, 386–390 (2014).
153. Waljee, A. K., Sauder, K., Zhang, Y., Zhu, J. & Higgins, P. D. R. External validation of a thiopurine monitoring algorithm on the SONIC clinical trial dataset. *Clin. Gastroenterol. Hepatol.* **16**, 449–451 (2018).
154. Waljee, A. K. et al. Predicting hospitalization and outpatient corticosteroid use in inflammatory bowel disease patients using machine learning. *Inflamm. Bowel Dis.* **24**, 45–53 (2018).
155. Wei, Z. et al. Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am. J. Hum. Genet.* **92**, 1008–1012 (2013).
156. Menti, E. 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**, 884–893 (2016).
157. Han, L. et al. A probabilistic pathway score (PROPS) for classification with applications to inflammatory bowel disease. *Bioinformatics* **34**, 985–993 (2018).
158. Hou, J. K. et al. Automated identification of surveillance colonoscopy in inflammatory bowel disease using natural language processing. *Dig. Dis. Sci.* **58**, 936–941 (2013).

Author contributions

L.P.-B., P.O., N.J. and G.N. researched data for the article. L.P.-B., P.O., S.D. and G.N. made substantial contributions to discussion of content for the article. L.P.-B., P.O., N.J. and G.N. wrote the article, and L.P.-B., P.O., S.D. and G.N. reviewed and edited the manuscript before submission.

Competing interests

P.O. has received financial support for research from Abbvie, Ferring and Takeda and lecture and consulting fees from Abbvie and Takeda. S.D. has received speaking, consultancy or advisory board member fees from Abbvie, Allergan, Biogen, Boehringer-Ingelheim, Celgene, Celltrion, Ferring, Hospira, Johnson and Johnson, Merck, MSD, Mundipharma, Pfizer, Sandoz, Takeda, TiGenix, UCB Pharma and Vifor. L.P.-B. has received consulting fees from Abbvie, Amgen, Biogaran, Boehringer-Ingelheim, Bristol-Myers Squibb, Celltrion, Ferring, Genentech, HAC Pharma, Hospira, Index Pharmaceuticals, Janssen, Lilly, Merck, Mitsubishi, Norgine, Pfizer, Pharmacosmos, Pilege, Sandoz, Takeda, Therakos, Tillotts, UCB Pharma and Vifor and lecture fees from Abbvie, Ferring, HAC Pharma, Janssen, Merck, Mitsubishi, Norgine, Takeda, Therakos, Tillotts and Vifor. N.J. and G.N. declare no competing interests.

Publisher's note

Reviewer information

Nature Reviews Gastroenterology & Hepatology thanks R. Panaccione, X. Roblin and S. Vavricka for their contribution to the peer review of this work.