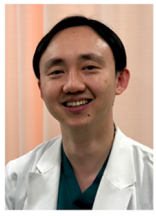
Gastroenterology 2022;162:1493–1506

Artificial Intelligence for Disease Assessment in Inflammatory Bowel Disease: How Will it Change Our Practice?



Ryan W. Stidham[1,2,*] Kento Takenaka[3,*]

[1]Division of Gastroenterology, Department of Internal Medicine, Michigan Medicine, Ann Arbor, Michigan; [2]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan; and [3]Department of Gastroenterology and Hepatology, Tokyo Medical and Dental University, Tokyo, Japan

Artificial intelligence (AI) has arrived and it will directly impact how we assess, monitor, and manage inflammatory bowel disease (IBD). Advances in the machine learning methodologies that power AI have produced astounding results for replicating expert judgment and predicting clinical outcomes, particularly in the analysis of imaging. This review will cover general concepts for AI in IBD, with descriptions of common machine learning methods, including decision trees and neural networks. Applications of AI in IBD will cover recent achievements in endoscopic image interpretation and scoring, new capabilities for cross-sectional image analysis, natural language processing for automated understanding of clinical text, and progress in AI-powered clinical decision support tools. In addition to detailing current evidence supporting the capabilities of AI for replicating expert clinical judgment, speculative commentary on how AI may advance concepts of disease activity assessment, care pathways, and pathophysiologic mechanisms of IBD will be addressed.

Keywords: Machine Learning; Crohn's Disease; Ulcerative Colitis; Inflammatory Bowel Disease; Artificial Intelligence.

An Introduction to Artificial Intelligence and Machine Learning

What is Artificial Intelligence?

Artificial intelligence (AI) is anticipated to transform the practice of medicine and inflammatory bowel disease (IBD) care by replicating the judgment of expert clinicians and discovering impactful insights through analyzing volumes of data too large and complex for humans to perceive. The concept of intelligence is broad and incorporates different capabilities. Intelligence requires the ability to acquire, store, and logically organize information. Intelligence also includes the ability to characterize information relationships with adjustment for situational context. Most critically, intelligence can apply known patterns and relationships between information and outcomes to infer or predict future events. Together, these facets of intelligence are the foundation for both clinical decision making and the creativity needed to yield new discoveries. Colloquially, AI refers to a synthetic or nonbiologic system that has some characteristics of intelligence. However, the adjective *artificial* is increasingly debated by computer scientists, neurologists, and philosophers as AI capabilities rapidly expand. In IBD, AI systems are proving to replicate complex measurements and judgments, showing promise for accurate and less biased disease measurement, predicting future clinical outcomes, and discovering new insights into the pathophysiology of disease.

The emergence of AI in IBD was made possible by the availability of high volumes of digitized medical data and the computational methods needed for complex pattern recognizing data analytics, collectively termed *machine learning* (ML). Physicians learn to make medical decisions by acquiring information, identifying patterns of information and observed outcomes, and continually learning from mistakes to improve their performance. In ML, input data are provided with annotations, labels, or classifications and can be clinical outcomes, expert measurements, or even

*Authors share co-first authorship.

Abbreviations used in this paper: **AI, artificial intelligence; ANN, artificial neural network; AUC, area under the curve; CAD, computer-aided diagnosis; CD, Crohn's disease; CTE, computed tomography enterography; IBD, inflammatory bowel disease; MES, Mayo Endoscopic Score; ML, machine learning; MRE, magnetic resonance enterography; NLP, natural language processing; RF, random forest; SVM, support vector machine; UC, ulcerative colitis; VCE, video capsule endoscopy.**

Most current article

0016-5085/$36.00
https://doi.org/10.1053/j.gastro.2021.12.238

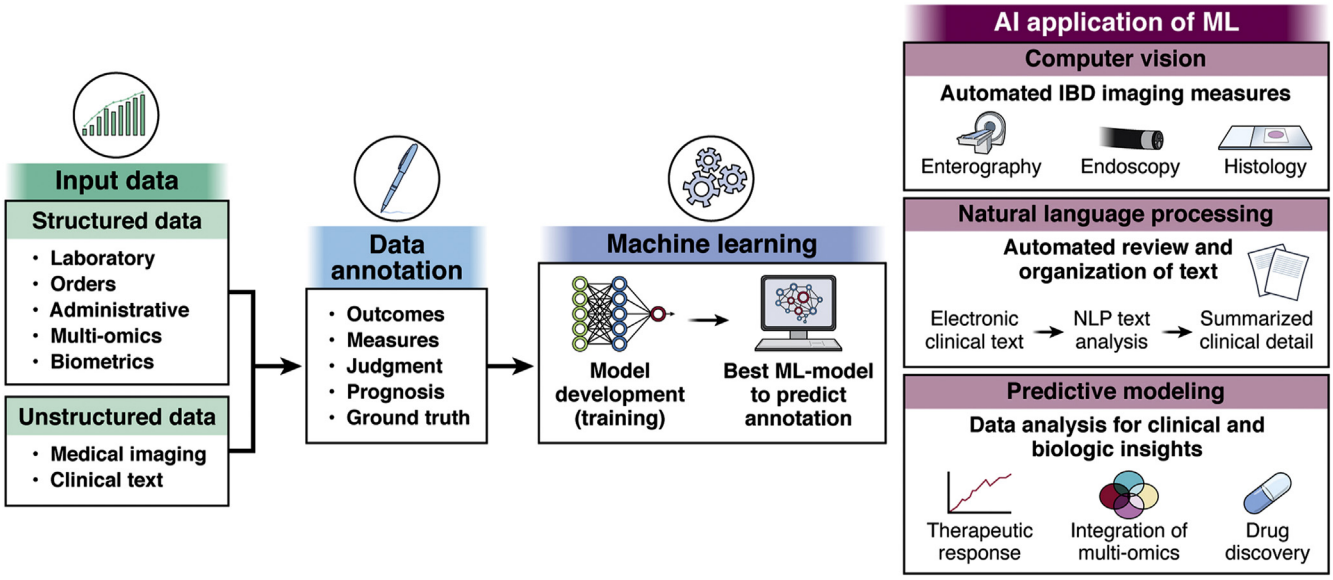


Figure 1. Summary of AI and related applications in IBD. AI's capabilities to mimic human understanding, judgment, and prediction rely on the availability of electronically stored data and the analytic capabilities to determine the data patterns associated with outcomes. Input data can be either structured or unstructured data types. Observations can be paired with annotations to denote the outcomes associated with a set of data variables. ML methods determine the relationships between sets of data and annotations to develop the best model for predicting an outcome (training). Model performance is calculated using a portion of the annotated dataset unseen during model development (testing). Applications of AI include computer vision for computational quantification and interpretation of an image, NLP for automated extraction of information from text, and predictive modeling for clinical outcomes and new scientific discoveries.

physiologic processes (Figure 1). ML methods quantify the relationships between the input data and the outcome as a model; this process is called *training*. During training, the strengths and patterns of relationships are iteratively modified to optimize the model parameters that best predict the outcome or other output. Model accuracy, reproducibility, and other performance metrics for predicting the outcome are tested on a new set of similar information. Supervised learning is when the output or outcome is provided during modeling. In unsupervised learning, an output or outcome is not provided, allowing data variables and population subgroups to be clustered into similar groupings. Unsupervised learning is commonly used for hypothesis generation and insight discovery.

ML advances are also dramatically improving the types of data that can be used in AI applications, specifically by making unstructured data available for analysis. Structured data are organized and readily analyzed with examples including spreadsheet formatted data like laboratory values, administrative claims, and biometric data. However, unstructured data sources have no inherent organization, despite being a repository for some of the most invaluable clinical data used in decision making, including medical imaging, endoscopic imaging, and text in clinical documents. ML methods can quantify and organize imaging and text as structured data better suited for computational analysis. These ML advances, specifically those allowing use of unstructured data, are most responsible for the dramatic expansion of AI capabilities in IBD and medicine at large.

Common Machine Learning Methodologies Used for Inflammatory Bowel Disease

Many ML methods are available, although support vector machines (SVMs), decision trees, and artificial neural networks (ANNs) are the most common ML used in IBD.[1,2] An SVM is an ML method for classification based on a set of values where each feature (variable) is considered a dimension of the dataset.[1] SVM calculate the boundary (also known as a plane) between all feature values to best split a dataset by a classification of interest (Figure 2*A*). Consider a fictitious example in which an investigator hypothesizes that colectomy risk in ulcerative colitis (UC) can be predicted by fecal calprotectin and length of colonic disease. Plotting calprotectin levels and disease length, an SVM classifier would identify the optimum boundary between the variables to best split the population studied by colectomy status. SVM decision splits can be straight lines or complex shapes that exist in hyperspace with *n*-dimensions. Criticisms of SVM include difficult interpretability, especially with nonlinear data, making implementation in clinical practice difficult.

Another common ML classification method is the decision tree, including a classification and regression tree or random forest (RF) ensemble methods.[3,4] Like SVM, decision trees are a supervised learning method that is used for a binary classification split. Consider an example in which an investigator wishes to predict future bowel resection with imaging and clinical variables in patients with a small bowel stenosis from Crohn's disease (CD). Using classification and regression tree (CART) methods, a

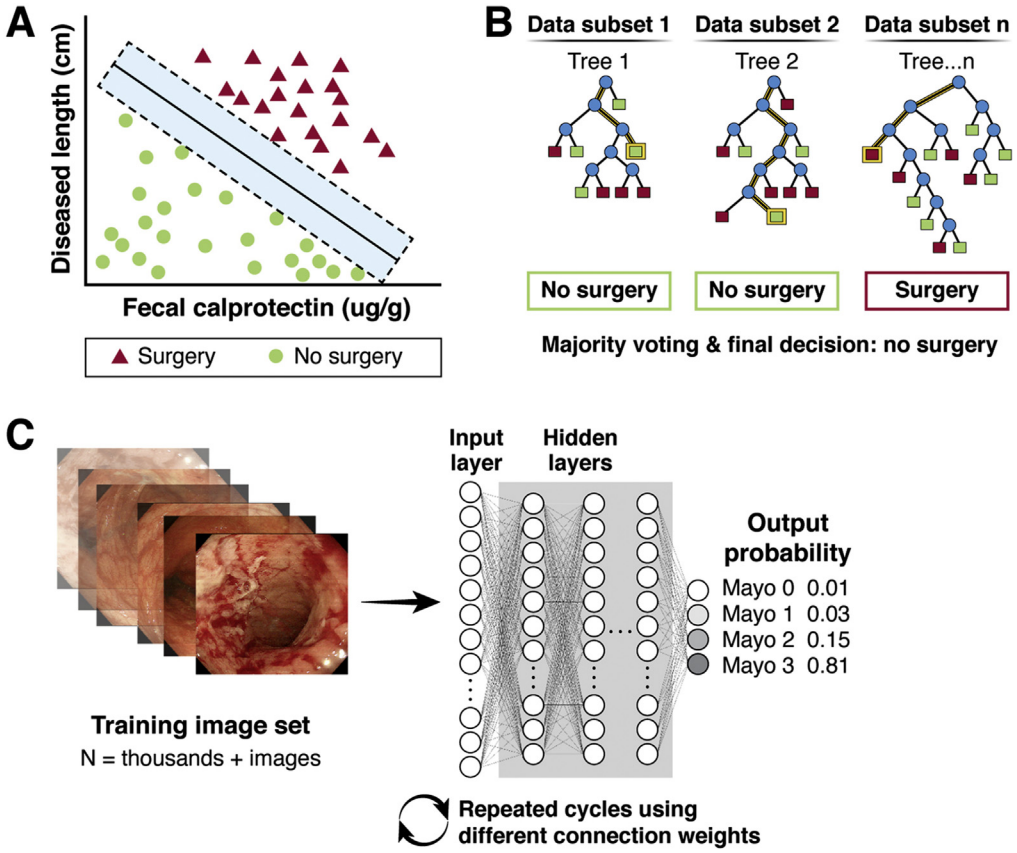


Figure 2. ML techniques commonly used in AI for IBD. (*A*) SVMs are designed to identify the boundary within a dataset that best separates 1 or more classes. In this example, 2 variables are plotted with the best boundary that optimally separates classes, here surgery vs no surgery. High-dimensional (multiple variables) and nonlinear data may also be used to find a classification boundary using SVM methods. (*B*) In RF methods, hundreds to thousands of decision trees are generated, and only a subset of the whole dataset and variables are used to optimize each decision tree for an event. Each tree provides a class prediction and the ensemble of trees vote to make a final prediction. Although individual trees in an RF are inferior to a single optimized decision tree, the ensemble voting approach better accounts for missing data and feature randomness. (*C*) ANNs can be used to analyze any type of data and perform particularly well for classifying imaging data. Image input layer features quantifying brightness, hue, correlation, and spatial data are expressed as values in nodes. The input layers are interconnected by several hidden layers. All nodes are weighted. When a weighted node value crosses an activation threshold, the downstream connected nodes are activated, terminating in 1 or more output nodes and generating a probability for each possible output. The essence of ANN training is the iterative adjustments to node weighting until the ANN prediction performance is maximized.

single best decision tree is constructed using the best cutoffs for each variable, positioning the most influential variables higher in the tree. Alternatively, RFs are an ensemble of many decision trees generated using only a portion of the features and a subset of the data (Figure 2*B*). Each tree provides a vote (or prediction) for the classification; the greatest number of votes is used to determine the final classification. The variation and diversity of features in RF outperform CART methods because RF better accounts for randomness, missing data, and error using the voting strategy. A key advantage of decision tree methods is interpretability, as clinicians often think in terms of decision splits.

ANNs are of particular importance in IBD, as their design is well-suited for image analysis (Figure 2*C*). ANNs loosely replicate biologic neural networks in the brain. For image classification, ANNs are designed as layers of nodes (neurons). In the setting of imaging ANNs, the input layer contains image features, such as brightness, hue, saturation, and homogeneity, as well as other unrecognizable descriptors, used to quantify images. Each input node value is then multiplied by a node weight. When the weighted input value crosses a threshold, connected nodes in downstream hidden layers are activated, similar to a synapse firing and activating connected neurons. Interconnected hidden layers ultimately terminate in an output node classifying the image or other data. Using a training set of thousands of images labeled with a classification of interest, an ANN learns by iteratively adjusting the weighting (or strength of relationship) between nodes to learn data characteristics associated with the expert-provided image annotation. This powerful model architecture is not restricted to images, but can also be used with other high-dimensional types of data, such as transcriptional, metabolomic, microbiome, or even clinical

Table 1. A Summary of Computer-Aided Characterization of Endoscopy for Ulcerative Colitis

First author	Year	Study design	No. of training samples	No. of test samples	Endoscopic imaging type	Outcome	Histology prediction
Maeda[17]	2019	Retrospective	Images from 87 patients	Images from 100 patients	Endocytoscopy (ultramagnifying endoscope) Still images	Predicting histologic remission (Geboes score <3.1)	Yes
Ozawa[8]	2019	Retrospective	Images from 841 patients	Images from 114 patients	Colonoscopy Still images	Identifying MES 0 or MES 0–1	No
Stidham[9]	2019	Retrospective	Images from 2778 patients	Images from 304 patients Videos from 30 patients	Colonoscopy Still images	Identifying individual MES grades of 0, 1, 2, and 3	No
Bossuyt[18]	2020	Prospective	Images from 29 patients	Images from 10 patients	Prototype endoscope Still images	Determining red density score, which correlated with endoscopic and histologic scores	Yes
Takenaka[14]	2020	Prospective	Images from 2012 patients	Images from 875 patients	Standard colonoscopy Still images	Determining UCEIS	Yes
Gottlieb[11]	2021	Prospective	Videos from 80% of 249 patients	Videos from 20% of 249 patients	Colonoscopy Complete video	Determining whole video MES and UCEIS	No
Yao[10]	2021	Prospective	Development set of 80 videos	51 videos of high-resolution sets; 264 videos from multicenter clinical trial sets	Colonoscopy Complete video	Determining whole video MES	No

UCEIS, Ulcerative Colitis Endoscopic Index of Severity; MES, Mayo Endoscopic Score.

data types for finding feature patterns associated with an outcome or event.

Using Artificial Intelligence for Analysis of Disease Activity Found on Imaging

Artificial Intelligence in Inflammatory Bowel Disease Endoscopy

Endoscopic impressions of IBD severity are the cornerstone of objective assessments in IBD and a key metric of therapeutic response.[5,6] Established scoring systems, even in the setting of central review, are limited by not only interobserver variation but also the qualitative nature of grading.[7] ANNs are showing promise for automating the interpretation of familiar endoscopic measures, particularity in UC (Table 1). Proof-of-concept studies for AI in UC have demonstrated the ability to replicate expert assessment of still endoscopic images in UC. A computer-aided diagnosis (CAD) system developed by Ozawa et al[8] using still colonoscopy images demonstrated excellent performance for distinguishing Mayo Endoscopic Score (MES) remission (MES 0, 1) from active disease (MES 2, 3; area under the receiver operating characteristic, 0.98). These results were reproduced by another group who developed a neural network MES model using more than 16,000 images from approximately 3000 patients with UC, with an excellent area under the receiver operating characteristic, sensitivity, and specificity of 0.970, 0.83, and 0.96, respectively, for separating endoscopic remission (MES 0, 1) from active disease (MES 2, 3).[9] In addition, agreement on the exact predicted MES grade was indistinguishable from the agreement between paired experienced reviewers ($\kappa = 0.84$ vs $\kappa = 0.86$), suggesting that automated systems performance is similar to experts for classifying UC severity on still images. Of course, still images are insufficient to provide a comprehensive impression of endoscopy, and methods are in development for whole endoscopic video interpretation. Challenges in analyzing entire endoscopic videos center around confounders, including bowel preparation, blurry images, biopsy-related injury, and other noise that must be distinguished from portions of the video suitable for grading. Using an array of sequential ANN classifiers, analysis of 315 complete endoscopic videos of both low- and high-resolution predicted MES with an agreement of 68% and 78%, respectively, compared with agreement of 82% among central readers.[10] Similar work was performed using full-length endoscopic video from the phase 2 trial of mirikizumab, where alternative methods had excellent agreement with central reader MES ($\kappa = 0.84$) and Ulcerative Colitis Endoscopic Index of Severity ($\kappa = 0.86$) scores.[11] Machines are demonstrating acceptable performance for

grading UC endoscopic severity and we should anticipate inclusion of automated assessments in clinical trials in the years to come.

Although replicating endoscopic measures in UC has immediate utility, AI is helping expand the capabilities of traditional endoscopy. Histologic activity, even in the setting of endoscopic remission, is a predictor of clinical outcomes in UC.[12,13] AI may detect subtle visual features on endoscopy, allowing histologic inference without biopsy. Takenaka et al[14] used 40,758 colonoscopy images paired with 6885 histologic interpretations to predict histologic remission (Geboes score <3) using only endoscopic images, with an accuracy of 93%.[15] In a prospective follow-up study, deep learning prediction of both endoscopic and histologic (deep) remission was associated with a significant reduction in hospitalization, colectomy, steroid use, and clinical relapse ($P < .001$).[16]

AI technologies may also help realize the benefits of new endoscopic visualization technologies designed to assist real-time histologic assessment during endoscopy. Endocytoscopy is an emerging technology in which a 520-fold ultramagnifying microscope is coupled with an endoscope, allowing real-time histologic assessment of colonic mucosa.[17] Recognizing the increasing relevance of histologic assessment for UC evaluation, Maeda et al[17] developed a CAD system to automatically provide real-time interpretation of endocytoscopy histologic activity in UC with a sensitivity, specificity, and accuracy of 74%, 97%, and 91%, respectively, using pathologist interpretation as a reference standard. As real-time capabilities for endoscopic histologic assessment become increasingly available, the AI technology providing reliable image interpretation is likely to transform our assessments of both disease activity and dysplasia detection and decision making. Bossuyt et al[18] are studying the ability to measure UC endoscopic activity using a red density score with a prototype endoscope designed to quantify image pixel color metrics and vessel pattern detection during colonoscopy.[18] The red density score is strongly correlated with the MES ($r = 0.76$, $P < .01$), Ulcerative Colitis Endoscopic Index of Severity ($r = 0.74$, $P < .01$), and the Robarts Histological Index ($r = 0.74$, $P < .01$). The red density score provides an increased dynamic range and may be found to be more objective and offer better discrimination between grades of disease activity than conventional endoscopic and histologic assessments alone. Together, these examples highlight how AI will help make sophisticated imaging technologies that are difficult to interpret more practical and feasible for clinical care.

Expect automated endoscopic disease activity scoring to soon be incorporated into therapeutic trials and eventually clinical practice. AI automated review of endoscopic video, either assisting expert reviewers or serving as a third digital reviewer, would improve the speed, efficiency, reproducibility, standardization, and cost of objective mucosal healing assessments. With the recent introduction of the first US Food and Drug Administration–approved endoscopic AI for detecting colonic polyps, a regulatory path for evaluating AI endoscopic technologies has been forged.[19] In addition, capabilities for providing a more granular analysis of endoscopic imaging is likely to produce new metrics of disease activity, potentially outperforming established measures like the MES, Ulcerative Colitis Endoscopic Index of Severity, and Simple Endoscopic Score for Crohn Disease.[20] AI endoscopic interpretation can aid remote telemedicine consultations, assist in education of trainees, and help monitor endoscopic quality.[21]

Although AI will change endoscopy forever, there are several barriers to overcome before implementing AI in routine clinical care. Who will serve as the ground truth reference for the endoscopic activity measures used to train AI CAD systems? As academics, nonprofits, and commercial enterprises aim to deploy AI systems for interpreting colonoscopy, attention to the characteristics, bias, and clinical conditions in the model training sets will be needed. Finally, AI systems will need to be explainable, producing evidence of the rationale for automated score predictions, as black box models will hinder trust and ultimately implementation.

Artificial Intelligence Applications in Capsule Endoscopy and Histology

The morphologic and anatomic variation typical of CD pose problems for current image analysis technologies using AI, resulting in limited success replicating common endoscopic scores, such as Simple Endoscopic Score for Crohn Disease and Crohn's Disease Endoscopic Index of Severity.[22] In contrast, current AI-based image classification is proving useful for detecting small bowel ulcerations using video capsule endoscopy (VCE).[23,24] Klang et al[25,26] have reported ANNs trained on VCE images can detect small bowel ulcerations with accuracy approximating 95%, and also can identify the presence of nonobstructive stenosis. Barash et al[27] developed an AI ulcer detection system capable of identifying different severity grades of CD ulcerations. Studying the potential clinical impact of using reliable AI methods for ulcer detection, Ding et al[28] found that automated methods reduced mean VCE review times from 96.6 minutes to 5.9 minutes with no difference in performance. Although the heterogeneity of CD presents challenges that will require further technologic developments, current methods may still prove useful in easing the time burden and improving sensitivity for reviewing VCE.

Similarly, automated quantitative description of histologic activity, which has progressed in oncology applications, will likely be seen in IBD.[29] Early work supporting the AI histologic analysis for IBD included a study of 287 pediatric patients with diagnostic accuracy of separating CD, UC, and non-IBD individuals of 76.9%.[30] In a study by Syed and colleagues,[31] a convolutional neural network was able to assess sets of duodenal biopsies distinguishing celiac disease, environmental enteropathy, and normal tissue with 93.4% accuracy. Interestingly, this neural network not only classified the underlying intestinal disease type, but highlighted the regions of the histologic images that most contributed to the diagnosis. Using a supervised neural network trained on histologic features of intestinal inflammation, Klein et al[32] reported the ability to use baseline intestinal biopsies to predict future phenotype, separating

B1 and B2 disease in 5 years' time with 70.5% accuracy. Although histologic interpretations are challenged by needing to account for variations in magnification level, depth of field, and the 3-dimensional space of the tissue assessed, AI-based histology for IBD will certainly progress. In addition, methods developed in other fields may provide new measures of histologic disease activity by reporting spatial relationships, cellular quantitation, and macroscopic architectural tissue features.[33]

Artificial Intelligence in Inflammatory Bowel Disease Cross-Sectional Imaging

Analogous to advancements in endoscopy, ML analytics are helping to improve IBD assessments using cross-sectional imaging. Magnetic resonance enterography (MRE) and computed tomography enterography (CTE) protocols are an essential companion to endoscopy for the diagnosis and longitudinal monitoring of CD disease phenotype, complications, and disease activity.[24,34] Cross-sectional imaging is substantially additive to the information collected by means of endoscopy.[35,36] Imaging features describing CD include continuous measures, such as bowel wall thickening, luminal narrowing, bowel dilation, and contrast enhancement, as well as qualitative features, including the presence or absence of mural stratification, mesenteric fat stranding, lymphadenopathy, and hypervascularity.[37,38] However, like endoscopy, variation between reviewers can limit measurement usefulness. Some measures have good interobserver agreement, such as wall thickness and edema in the terminal ileum, whereas other assessments, such as agreement on the presence of stenosis or penetration lesions, have proven to have higher variability, even when assessed by experienced radiologists.[39,40] Joint society guidelines and the development of several enterography disease scoring instruments have aimed to standardize definitions and operationalize disease quantification, with improvements over ad-hoc IBD activity reporting.[34,41,42] Yet, the variations among reviewer interpretation of imaging, differences in the availability of IBD expertise, and the time required to collect numerous measures highlight some of the key limitations in using cross-sectional imaging to its fullest potential.

Advancements in ML bowel segmentation methodologies for cross-sectional imaging over the last decade are powering both discrete disease activity measurements and inference of expert judgments.[43,44] Improvements in automated bowel segmentation are the key to automated extraction of standardized, reproducible CD activity measures using enterographies. Using 23 MREs in pediatric patients with CD, a neural network segmentation of the lumen, bowel wall, and background agreed with manually segmented bowel images in 75%, 81%, and 97% of cases, respectively.[45] Computer-assisted bowel wall thickness measurement in 53 MREs demonstrated lower measurement variance (0.46 mm^2 vs 2.90 mm^2; $P < .001$), greater intraclass correlation (0.88 vs 0.45; $P = .005$), and better spatial overlap (0.89 vs 0.72; $P < .0001$) compared with paired radiologists.[46] In 138 CTEs assessed by paired IBD radiologists, ML CTE analysis systems had measurement accuracy for bowel wall thickness ($P = .857$), maximum bowel dilation ($P = .557$), and minimum lumen diameter ($P = .596$) indistinguishable from radiologists.[47] ML also correctly inferred the presence of intestinal stenosis with accuracy of 84.4% (area under the receiver operating characteristic, 0.917) using anatomically extracted imaging data. Reliable AI extraction of conventional CD findings will aid in making objective, reproducible, and transparent measurements on IBD imaging more feasible in practice and research.

AI applications for enterography will also provide new perspectives on disease measurement. Conventional scoring of cross-sectional imaging typically evaluates disease features at a segment level, assuming homogeneity of features over the length of bowel. In a set of 207 CTE scans in patients with CD, a 3-dimensional neural network approach demonstrated the ability to map CD disease distribution at a resolution of 7.5 mm with an accuracy of 96.3%.[48] Bowel motility is an example of an imaging feature that confers information on CD activity and phenotype, but it is challenging to measure reliably.[49] Gollifer and colleagues[50] reported that AI-powered spatial ($P = .006$) and temporal ($P = .005$) deformation characteristics could quantify CD bowel motility properties, information that was correlated with Harvey-Bradshaw clinical indices. Finally, AI is already showing promise for making disease assessments that clinicians and radiologists have difficulty of providing. In a retrospective study of 167 patients with CD undergoing surgical resection of the ileum, investigators used a deep learning approach to extract more than 1400 individual radiomic features from the bowel to attempt prediction of the presence of moderate to severe intestinal fibrosis.[51] The radiomics model outperformed radiologists' impression on the presence of intestinal fibrosis after reviewing the same CTE studies, with an accuracy of 0.754 vs 0.590. Although radiomic models were imperfect, having a reported sensitivity and specificity for intestinal fibrosis of 0.811 and 0.679, respectively, this demonstrates an example when ML can use imaging to answer clinically relevant questions that are not easily assessed by human experts.

Potential Impact of Automated Analysis of Cross-Sectional Imaging on Inflammatory Bowel Disease Management

In the near term, AI in IBD cross-sectional imaging will likely be introduced as a new tool to help radiologists collect quantitative measures of disease activity. Although human verification and "sanity checks" of automatically generated measures will be needed, AI-powered tools will make collecting key IBD measures more feasible in clinical research and practice. In addition, expect a few (but not many) computer-generated imaging features that will prove clinically relevant in the care of IBD. Should professionals whose role centers on interpretating images begin preparing for a new career? Definitely not; they should prepare to be busier than ever. Information extraction from imaging and scoring of disease activity is a small role of radiologists,

pathologists, and gastroenterologists. ML is excellent at addressing specific questions and tasks, but lacks the flexibility, experience, and creativity to ask the right questions to provide comprehensive interpretations. The increase of information generated by AI imaging analysis will require more clinical expertise and spur an increase in the utilization of imaging.

Certainly more work is needed before IBD imaging analysis is adequate for use in practice. Bowel segmentation still requires significant correction due to difficulty negotiating penetrating complications, atypical abdominal anatomy, and ostomies. In addition, ML data analysis of current imaging will only take gastroenterology so far. Continued advancement in core imaging technologies are still needed, including improved ways to describe tissue, such as magnetization transfer MR imaging and stiffness imaging using bowel ultrasound.[52,53] However, as the amount and complexity of visual data increase, AI will be essential in coordinating new imaging information. AI systems will continue to improve, and the day when every CTE, MRE, and perhaps ultrasound will be automatically and systematically quantified for IBD activity is on the horizon.

Artificial Intelligence for Predicting Therapeutic and Clinical Outcomes

ML analysis of readily available conventional information sources, such as administrative claims data, diagnostic and procedural codes, and laboratory values, has shown promise for guiding decision making in IBD. ML methods, well-suited for pattern recognition, have been applied to predicting therapeutic response for a variety of IBD medications. Some of the earliest applications of AI included longitudinal therapeutic drug monitoring of thiopurines. Compared with traditional 6-TGN metabolite testing, ML algorithms using routinely collected laboratory values had superior performance predicting both clinical response (area under the curve [AUC], 0.59 vs 0.86) and objective biologic response (AUC 0.79 vs 0.49) among thiopurine users.[54,55] In addition, the thiopurine ML algorithm models had excellent performance predicting thiopurine shunting (AUC, 0.80) and medication noncompliance (AUC, 0.81). ML algorithm predictions of objective biologic response also conferred clinical benefits at 1 year, including reduction in surgical procedures (–0.5 events per year), hospitalizations (–1.5 events per year), and steroid prescriptions (–2.4 events per year). These prediction models were subsequently validated in the SONIC (Study of Biologic and Immunomodulator Naïve Patients in Crohn's Disease) clinical trial dataset and have been deployed at the University of Michigan as part of routine thiopurine monitoring.[56]

Similarly, ML methods have been used for prediction of response to biologic therapies, although the incremental benefits have not been dramatic. Using data from the GEMINI 1 and 2 studies of vedolizumab for UC, Waljee et al[57] used ensemble ML methods to integrate patient demographics, laboratory values, and vedolizumab drug levels to predict clinical remission at week 52, with a sensitivity and specificity of 0.76 and 0.71, respectively. The vedolizumab-response prediction model correctly predicted medication failure in 95.3% of patients using data from week 6 and 88.0% of cases using pretreatment data alone. RF ensemble methods were used to predict biologic response to ustekinumab in CD using the UNITI-1 and 2 studies.[58] Predictions of biologic remission at week 42 had a sensitivity and specificity of 0.79 and 0.67 using week 8 post-treatment data, although baseline pretreatment data had little predictive value (0.63 and 0.64, respectively). However, despite using high-quality clinical trial data, ML models offered minimal improvements compared with simply using fecal calprotectin and C-reactive protein to predict vedolizumab or ustekinumab response. ML sophistication is no guarantee of clinical utility, as advanced model architectures may be unable to overcome the limitations of the data collected.

ML analytics are also offering opportunities to analyze complex multi-omics data, identifying new IBD biomarkers and helping incorporate genetic, transcriptional, and microbiome data into clinical decisions. ML models using RNA expression levels from whole blood samples identified high- and low-risk profiles, which were strongly associated with the need for future therapeutic escalation in both CD (75% vs 35%) and UC (60% vs 20%).[59] In the same study, conventional high-risk clinical and endoscopic factors were unable to separate those needing therapeutic escalation over 18 months in CD ($P = .71$) and UC ($P = .36$). Studies aiming to predict CD recurrence after surgery reported that ML analysis of transcriptional data from resected ileal tissue successfully predicted disease recurrence at 1 year.[60] Acute severe UC outcome prediction has been aided by using neural networks to analyze expression data. Combining data from a pool of 3391 microRNA candidates and 5 clinical risk factors at admission for 47 patients with acute severe UC, the developed ML model separated medical responders from nonresponders with an accuracy of 97%; the microRNA-only model was 94% accurate.[61] ML is also assisting in the analytics of population-level genome-wide association studies (GWAS), proteomic and microbiome complex interactions to improve our understanding of disease and aid in biomarker development.[62–65]

How Could Artificial Intelligence Impact Clinical Decision Making in Real-World Inflammatory Bowel Disease Management?

ML capabilities can provide improved accuracy and additional insights into outcome prediction compared with traditional statistics. Expect next-generation clinical prediction models to incorporate automatically extracted imaging data, clinical, treatment, and laboratory data from electronic medical records, as well as emerging transcriptional, proteomic, and microbiome biomarkers. Modern ML outcome prediction models will continually learn and adapt, allowing iterative updates as new data become available.[66] Self-learning models may benefit from pooled data at a national level, but can also be customized to local data that better captures regional practice patterns and outcomes for improved prediction personalization. However, the clinical

utility of predictive models remains to be seen.[67] Expect dozens, if not hundreds, of IBD models predicting treatment outcomes, future relapse and disease progression. Inundated with AI predictions, gastroenterologists will need to increase our scrutiny of predictive model utility.[68] Seemingly impressive model performance statistics will need to be replaced with better assessments of clinical utility, such as decision curve analyses and other model net-benefit metrics.[69] The CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence) position statement provides excellent initial guidance on new considerations for developing AI models and evaluating their clinical utility.[70] If we are headed toward a future of starting our day with more automated electronic alerts and reminders from prediction models, we should ensure the AI suggestions are worth our precious attention.

Early Applications of Natural Language Processing for Inflammatory Bowel Diseases

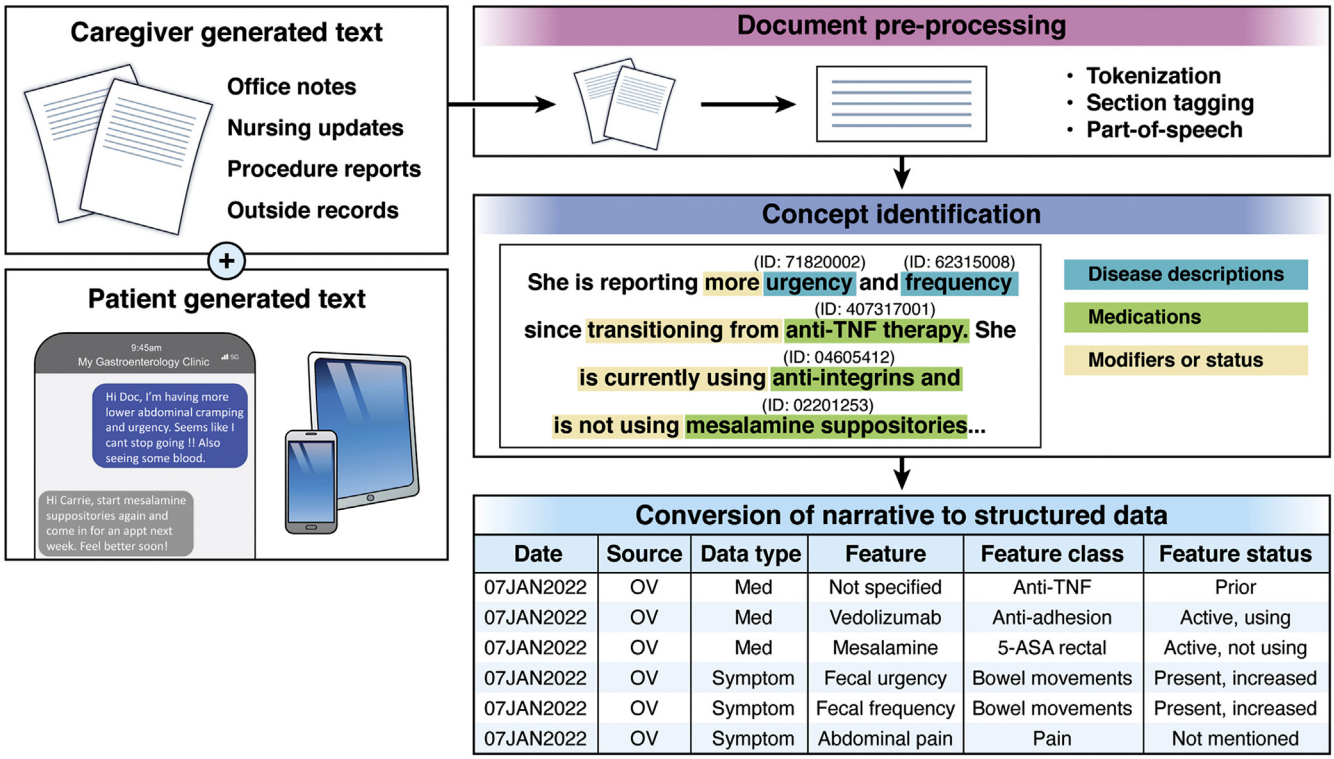
Decision making in IBD will always be strongly influenced by clinical features, symptoms, patient history, and experiences that are not captured by genomics, endoscopy, histology, or imaging. Natural language processing (NLP) is a field of AI designed to understand human text and can be used to automate collection of nuanced clinical data contained within electronic medical records.[71,72] NLP can help fill the information gaps in studies relying on administrative claims, diagnostic and procedure coding, and medication order records to characterize patients.[73] The development of NLP for IBD is in its infancy. However, progress supports expectations of a future state where patient-level phenotype, medication experience, and symptomatic course could be summarized in succinct data tables for analysis (Figure 3). Preliminary applications have highlighted NLP detecting 12% more patients with IBD than administrative diagnostic codes.[76] In addition, NLP analysis of clinical notes approximately doubled the detection of immunomodulator and biologic use and increased the detection of fistulizing (12% vs 36%) and stricturing (25% vs 40%) disease phenotypes in CD. NLP may be able to detect important IBD features, complications, and symptoms that are not routinely captured by administrative codes.

Beyond addressing the inconsistencies of administrative codes, NLP may be able to detect important IBD features, complications, and symptoms that are not coded routinely. Using a dataset of more than 1800 clinical notes, a pilot NLP pipeline was able to not only detect the mention of common extra-intestinal manifestations in clinical documents, but also could infer the degree of extra-intestinal manifestation activity with an overall sensitivity and specificity of 92.9% and 81.8%, respectively, compared with gastroenterologist interpretation.[77] In related work, another team used NLP to identify arthritis by automated review of clinical narratives, reporting a sensitivity and specificity of 0.83 and 0.92 compared with ICD-9 codes performance of 0.52 and 0.89, respectively.[78] NLP detection of arthritis eliminated the need for a difficult and laborious chart review and was used to show that vedolizumab users experienced arthralgia more commonly than anti–tumor necrosis factor users (46.1% vs 28.5%; $P = .002$). Shedding light on the inaccuracies of the information sources we rely upon for IBD phenotype and history, NLP may dramatically improve our ability to correctly describe IBD in individuals and populations.

NLP methods are also helping automatically understand patient-generated text. In a study of 1600 IBD-related social media posts, researchers using the ATLAS.ti NLP software package were able to automatically determine the theme and purpose of patient posts.[79] Sentiments, evolving concerns, and opinions of patients with IBD could be ascertained in real-time using adapted NLP methods that continuously analyze public social media posts. Researchers at the University of California-Los Angeles designed an NLP system with 95% accuracy classifying whether the reason for an incoming patient message was related to symptoms, medication problems, financial issues, procedures, laboratory results, or scheduling of appointments.[80] Real-time analysis of patient-generated text could aid in triage of messages and potentially provide suggested action or treatment. Considering these examples, it is not difficult to imagine NLP-powered digital companions assisting practitioners and patients in the management of IBD.

The Potential Impact of Natural Language Processing in the Management of Inflammatory Bowel Disease

NLP provides another example of using AI for complex information extraction. NLP may be the most difficult AI tool to develop, owing to the complexity of language, variation in documentation style between authors, and the unstructured nature of clinical narratives. Despite these limitations, the preliminary explorations discussed highlight the potential for NLP as a superior data source compared with administrative records for capturing disease history, medications, and potentially symptoms in IBD.[81] Expect advancements in NLP to provide clean structured datasets describing patients longitudinal IBD history. NLP can be performed passively as a background data analysis process likely to be incorporated into electronic medical record systems, potentially eliminating the need for patients, medical staff, or caregivers to complete surveys or long, structured questionnaires.[82–84] NLP-derived information could be used to quickly evaluate thousands of pages of outside records to determine history and interventions, responses to treatments, common symptoms, and to summarize prior investigations and studies in a comprehensive data summary. Finally, NLP interpretation of patient-generated narratives, either as e-mails, portal messages, or even text transcribed from telephone or video chats, could be used by chatbots and automated assistants for administrative tasks, such as medication refills and scheduling of routine health maintenance, and for interpreting symptoms and clinical disease status.



Conversion of narrative to structured data

Date	Source	Data type	Feature	Feature class	Feature status
07JAN2022	OV	Med	Not specified	Anti-TNF	Prior
07JAN2022	OV	Med	Vedolizumab	Anti-adhesion	Active, using
07JAN2022	OV	Med	Mesalamine	5-ASA rectal	Active, not using
07JAN2022	OV	Symptom	Fecal urgency	Bowel movements	Present, increased
07JAN2022	OV	Symptom	Fecal frequency	Bowel movements	Present, increased
07JAN2022	OV	Symptom	Abdominal pain	Pain	Not mentioned

Figure 3. NLP for extracting information from clinical text. Petabytes of text records, generated by health care providers and patients, present an enormous longitudinal narrative of IBD. However, extracting clinical information from highly variable natural speech is challenging. In NLP, documents undergo preprocessing to remove extraneous characters; partition documents into words, phrases, or sentences (tokenization); identify document sections; and detect parts of speech (verbs, nouns, adjectives). Next, disease features, descriptions, and other modifiers are detected and converted into uniform concepts using established concept libraries, including UMLS (Unified Medical Language System), MetaMap, and SNOMED-CT metathesauruses.[74,75] ML methods are used to associate concepts and infer meaning. The result is the conversion of an unstructured narrative into a structured data table, allowing more rapid retrieval of information by clinicians, also powering AI models for decision support systems.

Conclusions

AI will impact nearly every aspect of IBD—how we assess disease, make treatment decisions, discovery of new biomarkers and medications, and even our communication with patients. However, despite enthusiasm for an AI-enhanced IBD practice, we must prepare for new responsibilities and challenges. All IBD providers will need to become informed consumers of AI. IBD education must remain ahead of the AI implementation curve. Education of future IBD practitioners should include core competencies for understanding ML, evaluating AI performance, and best practices for quality assurance in using automated analysis and prediction models.[85,86] Education must also include awareness and attention to potential demographic and racial bias that can be introduced by AI.[87–89] Ideally, beginning today, IBD education will integrate applied predictive models, image analyses, and digital personalization applications in tandem with traditional clinical experiences to produce our first generation of caregivers assisted by AI.

Conversations regarding the yet uncharted medical-legal landscape presented by AI should begin soon. Conflicts between AI and clinicians will definitely arise and a framework of the responsibilities for caregivers and the AI guarantors should be devised before implementation.[90,91] Transparent discussions and practices regarding data privacy, ownership, and access privileges for research, private companies, payors, and governments will be important ethical considerations. Bundled with the promise of benefits for IBD, these additional considerations must be addressed for the most harmonious integration of AI into clinical practice.

Although how AI will be integrated in clinical care remains to be seen, the immediate influence of AI in IBD will center around the collection, extraction, and organization of clinical information (Figure 4). Expect with near certainty that disease grading, particularly endoscopic, histologic, and eventually cross-sectional imaging assessments, will be replaced by instantaneous, standardized, and high-reliability AI systems. Prepare to surrender IBD disease activity assessment to the machines. Furthermore, NLP improvements in document understanding will reduce the time needed to find and organize a patient's clinical information, presently scattered across thousands of pages of electronic documents. The art of the chart review may become a welcome relic by the mid 21st century.

Will the physician become irrelevant? Although information extraction will increasingly be in the purview of

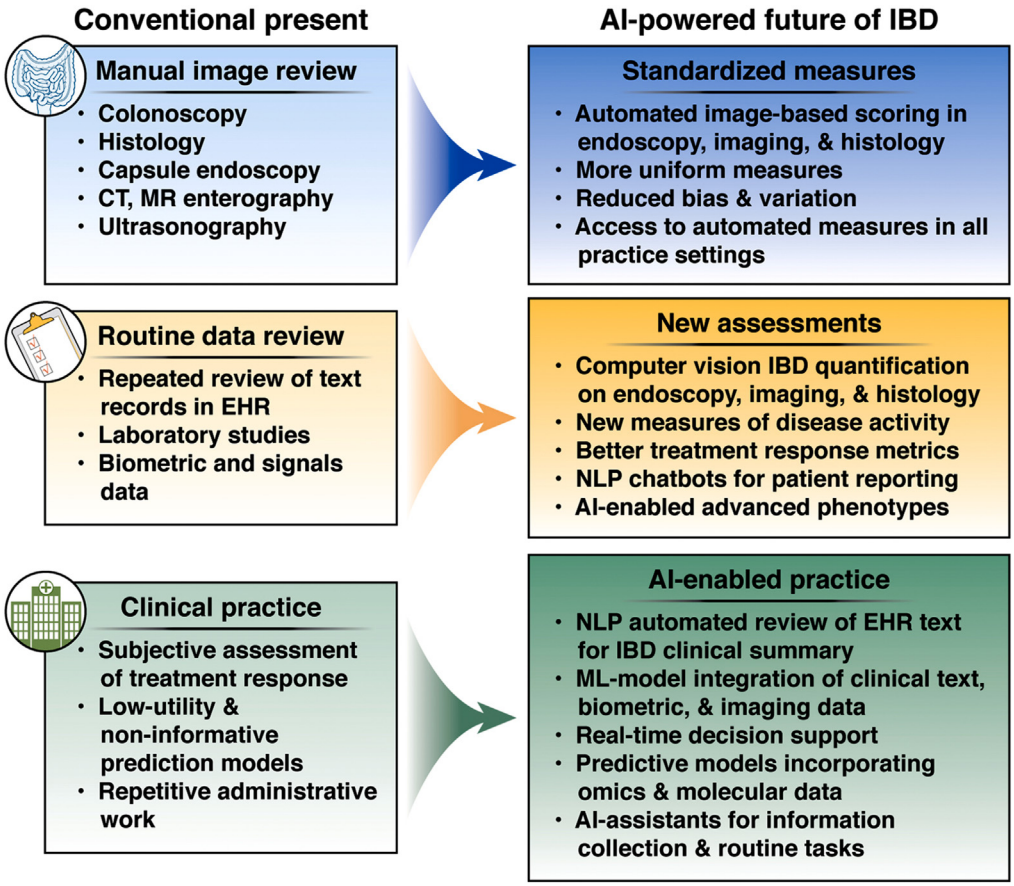


Figure 4. Expectations of AI use in IBD clinical care. Near-term applications of AI into IBD care will most likely involve standardization measurements with the benefits of automation. Conventional interpretation and more tedious disease activity measures on endoscopy, cross-sectional imaging, and histology will be automatically extracted by customized ML methods with more reproducibility, objectivity, and at high speed. Expect early applications to serve as a digital assistant, with measurements verified by experienced clinicians but, in time, AI measurements will prove reliability without the need for human review of results. Expanded disease quantification will power more granular and specific measures of disease activity and potentially even new phenotypes for UC and CD. Finally, AI systems will power better clinical decision support systems, integrating current and historical available data, including patient generated data, to assist in routine IBD management.

machines, clinicians will continue to be the decision makers. The IBD physician of the near future will cease collecting information that will instead be provided by AI assistants, liberating much needed time for more critical thinking, data analysis, and education of our patients. Direct care delivery by AI is likely decades away. AI will be challenged with understanding routine nuances in IBD care, such as the reasons why a patient hesitates to use immunosuppression, why a colonoscopy may be needed outside of the standard surveillance window, or when to continue treatment despite a mild adverse effect. Even in the information extraction domain, AI systems are likely to be incapable of collecting the wide variety of emotional and psychosocial information that physicians intuitively comprehend and use in decision making. AI will need supervision for years to gain our confidence, as it remains inexperienced, only knows what it is trained, only can know the information it has access to, and is currently inflexible to situation and clinical context.

In the coming decade, the quality of clinical, physiologic, and molecular disease descriptions enhanced by AI will enable the next era of personalized medicine. Practitioners will be able to quickly examine trends in disease activity, symptoms, and outcomes at individual and population levels. New AI-generated measures of clinical imaging could prove superior to expert impressions of disease activity impossible to be detected by the unaided human eye. Time will tell whether AI-powered advances will improve the utility of clinical decision support tools over current care pathways. However, there is no question that AI will be integrated into our clinical practice as an unbiased judge, an assistant, and possibly as a surrogate. We are at the very beginning of the digital revolution for IBD and we have complete control of how AI is implemented in the 21st century IBD care.

References

1. PLoS Medicine Editors. Advancing the beneficial use of machine learning in health care and medicine: toward a community understanding. In: Nevin L, ed. PLoS Med 2018;15(11):e1002708.

2. de Bruijne M. Machine learning approaches in medical image analysis: from detection to diagnosis. Med Image Anal 2016;33:94–97.
3. Che D, Liu Q, Rasheed K, et al. Decision tree and ensemble learning algorithms with their applications in bioinformatics. In: Arabnia HR, Tran QN, eds. *Software Tools and Algorithms for Biological Systems*. Springer, 2011:191–199.
4. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. J Comp Syst Sci 1997;55:119–139.
5. Turner D, Ricciuto A, Lewis A, et al. STRIDE-II: an update on the Selecting Therapeutic Targets in Inflammatory Bowel Disease (STRIDE) Initiative of the International Organization for the Study of IBD (IOIBD): determining therapeutic goals for treat-to-target strategies in IBD. Gastroenterology 2021;160:1570–1583.
6. Feagan BG, Sandborn WJ, D'Haens G, et al. The role of centralized reading of endoscopy in a randomized controlled trial of mesalamine for ulcerative colitis. Gastroenterology 2013;145:149–157.e2.
7. Osada T, Ohkusa T, Yokoyama T, et al. Comparison of several activity indices for the evaluation of endoscopic activity in UC: inter- and intraobserver consistency. Inflamm Bowel Dis 2010;16:192–197.
8. Ozawa T, Ishihara S, Fujishiro M, et al. Novel computer-assisted diagnosis system for endoscopic disease activity in patients with ulcerative colitis. Gastrointest Endosc 2019;89:416–421.e1.
9. 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.
10. Yao H, Najarian K, Gryak J, et al. Fully automated endoscopic disease activity assessment in ulcerative colitis. Gastrointest Endosc 2021;93:728–736.e1.
11. Gottlieb K, Requa J, Karnes W, et al. Central reading of ulcerative colitis clinical trial videos using neural networks. Gastroenterology 2021;160:710–719.e2.
12. Kaneshiro M, Takenaka K, Suzuki K, et al. Pancolonic endoscopic and histologic evaluation for relapse prediction in patients with ulcerative colitis in clinical remission. Aliment Pharmacol Ther 2021;53:900–907.
13. Cushing KC, Tan W, Alpers DH, et al. Complete histologic normalisation is associated with reduced risk of relapse among patients with ulcerative colitis in complete endoscopic remission. Aliment Pharmacol Ther 2020; 51:347–355.
14. Takenaka K, Ohtsuka K, Fujii T, et al. Development and validation of a deep neural network for accurate evaluation of endoscopic images from patients with ulcerative colitis. Gastroenterology 2020;158:2150–2157.
15. Geboes K, Riddell R, Ost A, et al. A reproducible grading scale for histological assessment of inflammation in ulcerative colitis. Gut 2000;47:404–409.
16. Takenaka K, Ohtsuka K, Fujii T, et al. Deep neural network accurately predicts prognosis of ulcerative colitis using endoscopic images. Gastroenterology 2021; 160:2175–2177.e3.
17. Maeda Y, Kudo S-E, Mori Y, et al. Fully automated diagnostic system with artificial intelligence using endocytoscopy to identify the presence of histologic inflammation associated with ulcerative colitis (with video). Gastrointest Endosc 2019;89:408–415.
18. Bossuyt P, Vermeire S, Bisschops R. Scoring endoscopic disease activity in IBD: artificial intelligence sees more and better than we do. Gut 2020;69:788–789.
19. Repici A, Badalamenti M, Maselli R, et al. Efficacy of real-time computer-aided detection of colorectal neoplasia in a randomized trial. Gastroenterology 2020; 159:512–520.e7.
20. Gottlieb K, Daperno M, Usiskin K, et al. Endoscopy and central reading in inflammatory bowel disease clinical trials: achievements, challenges and future developments. Gut 2021;70:418–426.
21. Cross RK, Langenberg P, Regueiro M, et al. A randomized controlled trial of TELEmedicine for Patients with Inflammatory Bowel Disease (TELE-IBD). Am J Gastroenterol 2019;114:472–482.
22. Sipponen T, Nuutinen H, Turunen U, et al. Endoscopic evaluation of Crohn's disease activity: comparison of the CDEIS and the SES-CD. Inflamm Bowel Dis 2010; 16:2131–2136.
23. Enns RA, Hookey L, Armstrong D, et al. Clinical practice guidelines for the use of video capsule endoscopy. Gastroenterology 2017 2017;152:497–514.
24. 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 Crohns Colitis 2019;13:273–284.
25. Klang E, Barash Y, Margalit RY, et al. Deep learning algorithms for automated detection of Crohn's disease ulcers by video capsule endoscopy. Gastrointest Endosc 2020;91:606–613.e2.
26. Klang E, Grinman A, Soffer S, et al. Automated detection of Crohn's disease intestinal strictures on capsule endoscopy images using deep neural networks. J Crohns Colitis 2021;15:749–756.
27. Barash Y, Azaria L, Soffer S, et al. Ulcer severity grading in video capsule images of patients with Crohn's disease: an ordinal neural network solution. Gastrointest Endosc 2021;93:187–192.
28. Ding Z, Shi H, Zhang H, et al. Gastroenterologist-level identification of small-bowel diseases and normal variants by capsule endoscopy using a deep-learning model. Gastroenterology 2019;157:1044–1054.e5.
29. Bandi P, Geessink O, Manson Q, et al. From detection of individual metastases to classification of lymph node status at the patient level: the CAMELYON17 Challenge. IEEE Trans Med Imaging 2019;38:550–560.
30. Mossotto E, Ashton JJ, Coelho T, et al. Classification of paediatric inflammatory bowel disease using machine learning. Sci Rep 2017;7. 2427–2410.
31. Syed S, Al-Boni M, Khan MN, et al. Assessment of machine learning detection of environmental enteropathy and celiac disease in children. JAMA Netw Open 2019;2: e195822–e195822.

32. Klein A, Mazor Y, Karban A, et al. Early histological findings may predict the clinical phenotype in Crohn's colitis. United European Gastroenterol J 2017; 5:694–701.
33. Pradhan P, Meyer T, Vieth M, et al. Semantic segmentation of non-linear multimodal images for disease grading of inflammatory bowel disease: a SegNet-based application. Available at: https://www.scitepress.org/Papers/2019/73140/73140.pdf. Accessed June 5, 2021.
34. Bruining DH, Zimmermann EM, Loftus EV, et al. Consensus recommendations for evaluation, interpretation, and utilization of computed tomography and magnetic resonance enterography in patients with small bowel Crohn's disease. Gastroenterology 2018; 154:1172–1194.
35. Lichtenstein GR, Loftus EV, Isaacs KL, et al. ACG Clinical Guideline: Management of Crohn's Disease in Adults. Am J Gastroenterol 2018;113:481–517.
36. Lewis JD, Rutgeerts P, Feagan BG, et al. Correlation of stool frequency and abdominal pain measures with Simple Endoscopic Score for Crohn's Disease. Inflamm Bowel Dis 2020;26:304–313.
37. Al-Hawary MM, Kaza RK, Platt JF. CT enterography: concepts and advances in Crohn's disease imaging. Radiol Clin North Am 2013;51:1–16.
38. Al-Hawary MM, Zimmermann EM, Hussain HK. MR imaging of the small bowel in Crohn disease. Magn Reson Imaging Clin N Am 2014;22:13–22.
39. Jairath V, Ordás I, Zou G, et al. Reliability of measuring ileo-colonic disease activity in Crohn's disease by magnetic resonance enterography. Inflamm Bowel Dis 2018;24:440–449.
40. Rees MA, Dillman JR, Anton CG, et al. Inter-radiologist agreement using Society of Abdominal Radiology-American Gastroenterological Association (SAR-AGA) consensus nomenclature for reporting CT and MR enterography in children and young adults with small bowel Crohn disease. Abdom Radiol (N Y) 2019; 44:391–397.
41. Ordás I, Rimola J, Alfaro I, et al. Development and validation of a simplified magnetic resonance index of activity for Crohn's disease. Gastroenterology 2019; 157:432–439.e1.
42. Rimola J, Alvarez-Cofiño A, Pérez-Jeldres T, et al. Comparison of three magnetic resonance enterography indices for grading activity in Crohn's disease. J Gastroenterol 2017;52:585–593.
43. Mahapatra D, Schüffler PJ, Tielbeek JAW, et al. Combining multiple expert annotations using semi-supervised learning and graph cuts for Crohn's Disease segmentation. In: Yoshida Y, Näppi JJ, Saini S, eds. Abdominal Imaging: Computational and Clinical Applications. Springer, 2014:139–147.
44. Mahapatra D, Vos FM, Buhmann JM. Active learning based segmentation of Crohn's disease from abdominal MRI. Comput Methods Programs Biomed 2016; 128:75–85.
45. Lamash Y, Kurugol S, Freiman M, et al. Curved planar reformatting and convolutional neural network-based segmentation of the small bowel for visualization and quantitative assessment of pediatric Crohn's disease from MRI. J Magn Reson Imaging 2019;49:1565–1576.
46. Naziroglu RE, Puylaert CAJ, Tielbeek JAW, et al. Semi-automatic bowel wall thickness measurements on MR enterography in patients with Crohn's disease. Br J Radiol 2017;90:20160654.
47. 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 2019;11:274.
48. Enchakalody BE, Henderson B, Wang SC, et al. Machine learning methods to predict presence of intestine damage in patients with Crohn's disease. In: Hahn HK, Mazurowski MA, eds. *Medical Imaging 2020: Computer-Aided Diagnosis*11314. International Society for Optics and Photonics, 2020:1131436.
49. Menys A, Puylaert C, Tutein Nolthenius CE, et al. quantified terminal ileal motility during MR enterography as a biomarker of Crohn disease activity: prospective multi-institution study. Radiology 2018;289:428–435.
50. Gollifer RM, Menys A, Plumb A, et al. Automated versus subjective assessment of spatial and temporal MRI small bowel motility in Crohn's disease. Clin Radiol 2019;74. 814.e9–814.e19.
51. Li X, Liang D, Meng J, et al. Development and validation of a novel computed-tomography enterography radiomic approach for characterization of intestinal fibrosis in Crohn's disease. Gastroenterology 2021; 160:2303–2316.e11.
52. Adler J, Swanson SD, Schmiedlin-Ren P, et al. Magnetization transfer helps detect intestinal fibrosis in an animal model of Crohn disease. Radiology 2011; 259:127–135.
53. Stidham RW, Higgins PD. Imaging of intestinal fibrosis: current challenges and future methods. United European Gastroenterol J 2016;4:515–522.
54. 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;8:143–150.
55. Waljee AK, Sauder K, Patel A, et al. Machine learning algorithms for objective remission and clinical outcomes with thiopurines. J Crohns Colitis 2017;11:801–810.
56. Waljee AK, Sauder K, Zhang Y, et al. External validation of a thiopurine monitoring algorithm on the SONIC clinical trial dataset. Clin Gastroenterol Hepatol 2018; 16:449–451.
57. Waljee AK, Liu B, Sauder K, et al. Predicting corticosteroid-free biologic remission with vedolizumab in Crohn's disease. Inflamm Bowel Dis 2018; 24:1185–1192.
58. Waljee AK, Wallace BI, Cohen-Mekelburg S, et al. Development and validation of machine learning models in prediction of remission in patients with moderate to severe Crohn disease. JAMA Netw Open 2019;2: e193721.
59. Biasci D, Lee JC, Noor NM, et al. A blood-based prognostic biomarker in IBD. Gut 2019;68:1386–1395.
60. 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;25:180–193.
61. Morilla I, Uzzan M, Laharie D, et al. Colonic MicroRNA profiles, identified by a deep learning algorithm, that predict responses to therapy of patients with acute severe ulcerative colitis. Clin Gastroenterol Hepatol 2019; 17:905–913.
62. 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;23:1516–1523.
63. 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.
64. Andrighetti T, Bohar B, Lemke N, et al. MicrobioLink: an integrated computational pipeline to infer functional effects of microbiome-host interactions. Cells 2020; 9(5):1278.
65. Ungaro RC, Hu L, Ji J, et al. Machine learning identifies novel blood protein predictors of penetrating and stricturing complications in newly diagnosed paediatric Crohn's disease. Aliment Pharmacol Ther 2021; 53:281–290.
66. Kaul V, Enslin S, Gross SA. History of artificial intelligence in medicine. Gastrointest Endosc 2020; 92:807–812.
67. Shah NH, Milstein A, Bagley SC. Making machine learning models clinically useful. JAMA 2019; 322:1351–1352.
68. Wynants L, van Smeden M, McLernon DJ, et al. Three myths about risk thresholds for prediction models. BMC Med 2019;17:192–197.
69. Vickers AJ, Van Calster B, Steyerberg EW. Net benefit approaches to the evaluation of prediction models, molecular markers, and diagnostic tests. BMJ 2016;352:i6.
70. Liu X, Rivera SC, Moher D, et al. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. BMJ 2020; 370:m3164.
71. Masanz J, Pakhomov SV, Xu H, et al. Open Source Clinical NLP - more than any single system. AMIA Jt Summits Transl Sci Proc 2014;2014:76–82.
72. Soysal E, Wang J, Jiang M, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc 2018; 25:331–336.
73. Van Vleck TT, Chan L, Coca SG, et al. Augmented intelligence with natural language processing applied to electronic health records for identifying patients with non-alcoholic fatty liver disease at risk for disease progression. Int J Med Inform 2019;129:334–341.
74. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17:229–236.
75. Kang T, Perotte A, Tang Y, et al. UMLS-based data augmentation for natural language processing of clinical research literature. J Am Med Inform Assoc 2021; 28:812–823.
76. 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: a novel informatics approach. Inflamm Bowel Dis 2013;19:1411–1420.
77. Stidham RW, Yu D, Lahiri S, Vydiswaran V. P311 Detection and characterisation of extra-intestinal manifestations of IBD in clinical office notes using natural language processing. J Crohn's Colitis 2020;14(Suppl 1): S309–S310.
78. Cai T, Lin T-C, Bond A, et al. The association between arthralgia and vedolizumab using natural language processing. Inflamm Bowel Dis 2018;24:2242–2246.
79. Martinez B, Dailey F, Almario CV, 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 2017;23:1057–1064.
80. Zand A, Sharma A, Stokes Z, et al. An exploration into the use of a chatbot for patients with inflammatory bowel diseases: retrospective cohort study. J Med Internet Res 2020;22:e15589.
81. Liao KP, Cai T, Savova GK, et al. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. BMJ 2015;350:h1885.
82. Imler TD, Morea J, Kahi C, et al. Multi-center colonoscopy quality measurement utilizing natural language processing. Am J Gastroenterol 2015; 110:543–552.
83. Imler TD, Sherman S, Imperiale TF, et al. Provider-specific quality measurement for ERCP using natural language processing. Gastrointest Endosc 2018; 87:164–173.e2.
84. Kurowski JA, Milinovich A, Ji X, et al. Differences in biologic utilization and surgery rates in pediatric and adult Crohn's disease: results from a large electronic medical record-derived cohort. Inflamm Bowel Dis 2021; 27:1035–1044.
85. Embi PJ. Algorithmovigilance-advancing methods to analyze and monitor artificial intelligence-driven health care for effectiveness and equity. JAMA Netw Open 2021;4. e214622–e214622.
86. Park Y, Jackson GP, Foreman MA, et al. Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA Open 2020;3:326–331.
87. Obermeyer Z, Powers B, Vogeli C, et al. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019;366:447–453.
88. Bailey ZD, Feldman JM, Bassett MT. How structural racism works - racist policies as a root cause of U.S. racial health inequities. N Engl J Med 2021; 384:768–773.
89. Tat E, Bhatt DL, Rabbat MG. Addressing bias: artificial intelligence in cardiovascular medicine. Lancet Digital Health 2020;2:e635–e636.

90. Holm S, Stanton C, Bartlett B. A new argument for no-fault compensation in health care: the introduction of artificial intelligence systems. Health Care Anal 2021;29:171–188.

91. Jaremko JL, Azar M, Bromwich R, et al. Canadian Association of Radiologists White Paper on Ethical and Legal Issues Related to Artificial Intelligence in Radiology. Can Assoc Radiol J 2019;70:107–118.

Received May 21, 2021. Accepted December 6, 2021.

Correspondence
Address correspondence to: Ryan Stidham, MD, MS, Department of Computational Medicine and Bioinformatics, University of Michigan, 3912 Taubman Center, 1500 East Medical Center Drive, Ann Arbor, Michigan 48109. e-mail: ryanstid@med.umich.edu.

CRediT Authorship Contributions
Conceptualization, writing and review of the manuscript were equally performed by RWS and KT.

Conflicts of interest
The authors disclose the following: Ryan W. Stidham has served as a consultant or on advisory boards for AbbVie, Janssen, Takeda, Gilead, Eli Lilly, Exact Sciences, Evergreen Pharmaceuticals, and CorEvitas; and holds intellectual property on cross-sectional imaging and endoscopic analysis technologies licensed by the University of Michigan to AMI, Inc. Kento Takenaka has served as a consultant for AbbVie, Janssen, and Takeda.

Funding
Funding support to Ryan W. Stidham was provided by the National Institute of Diabetes and Digestive and Kidney Diseases (R01-DK124779).