DR. TYLER M BERZIN (Orcid ID : 0000-0002-4364-6210)

Article type : Review

Charting a Path Forward for Clinical Research in Artificial Intelligence and Gastroenterology

Jeremy R. Glissen Brown MD,[1] Akbar K. Waljee MD,[2] Yuichi Mori MD, PhD, [3, 4] Prateek Sharma MD[5,6] and Tyler M. Berzin MD, MS, FASGE[1]

1. Center for Advanced Endoscopy, Division of Gastroenterology, Beth Israel Deaconess Medical Center and Harvard Medical School, Boston, Massachusetts
2. Division of Gastroenterology, University of Michigan Health System, University of Michigan, Ann Arbor, Michigan
3. Digestive Disease Center, Showa University Northern Yokohama Hospital, Yokohama, Japan
4. Clinical Effectiveness Research Group, Institute of Health and Society, University of Oslo, Oslo, Norway
5. Department of Gastroenterology and Hepatology, University of Kansas Medical Center, Kansas City, KS, USA
6. Department of Gastroenterology, Kansas City VA Medical Center, Kansas City, KS, USA

Corresponding Author:

Jeremy R. Glissen Brown MD

Center for Advanced Endoscopy, Division of Gastroenterology and Hepatology
Beth Israel Deaconess Medical Center and Harvard Medical School
330 Brookline Avenue
Boston, MA 02130
jglissen@bidmc.harvard.edu

Keywords: SPIRIT-AI, CONSORT-AI, Deep learning, Machine learning, Guidelines

Conflict of Interests:
Author T.M.B. has received consultant fees from Wision AI, Fujifilm and Medtronic. Author Y.M. has received consultant fees and speaking honoraria from Olympus Corp. Author P.S. has received grant support from Ironwood, Erbe, Docbot, Cosmo pharmaceuticals and CDx labs. Author P.S. has received consultant fees from Medtronic, Olympus, Boston Scientific, Fujifilm and Lucid. Authors J.R.G.B and A.K.B. have no conflicts of interest.

Author Contributions:
Jeremy Glissen Brown, Tyler Berzin and Akbar Waljee were involved in concept design. All authors were involved in interpretation of studies involved. Jeremy Glissen Brown drafted the initial manuscript. Akbar Waljee, Yuichi Mori, Tyler Berzin and Prateek Sharma were involved in manuscript review and critical revisions. All authors read and approved the final manuscript.

Word Count: 6183

Abbreviations:
AI – Artificial Intelligence
ML – Machine learning
VCE – Video Capsule Endoscopy
CADe – Computer Aided Detection
IBD – Inflammatory Bowel Disease
IBS – Irritable Bowel Syndrome

CADx – Computer Aided Diagnosis

ADR – Adenoma Detection Rate

RR – Relative Risk

CONSORT – Consolidated Standards of Reporting Trials

SPIRIT – Standard Protocol Items: Recommendations for Interventional Trials

ESGE – European Society of Gastrointestinal Endoscopy

ASGE – American Society for Gastrointestinal Endoscopy

AuROC – Area under the receiver operating characteristic curve

RF – Random Forest

TRIPOD – Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis

PROBAST – Prediction model risk of bias assessment tool

TREE – Transparent, reproducible, ethical and effective

Gastroenterology has been an early leader in bridging the gap between artificial intelligence model development and clinical trial validation and in recent years we have seen the publication of several randomized clinical trials examining the role of artificial intelligence in gastroenterology. As AI applications for clinical medicine advance rapidly, there is a clear need for guidance surrounding AI-specific study design, evaluation, comparison, analysis and reporting of results. Several initiatives are in the publication or pre-publication phase including AI-specific amendments to minimum reporting guidelines for clinical trials, society task force initiatives aimed at priority use cases and research priorities and minimum reporting guidelines that guide the reporting of clinical prediction models. In this paper we examine applications of AI in clinical trials and discuss elements of newly published AI-specific extensions to the Consolidated Standards of Reporting Trials (CONSORT) and Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) statements that guide clinical trial reporting and development. We will then review AI-applications at the pre-trial level in both endoscopy and other subfields of gastroenterology and explore areas where further guidance is needed to supplement the current guidance available at the pre-trial level.

INTRODUCTION

Over the past decade, artificial intelligence (AI) has captured the popular imagination and has been the object of intense media and commercial focus due in large part to recent applications in facial recognition, natural language processing, autonomous driving and medical imaging. The field of machine learning (ML) – a set of computational methods that involves using mathematical models to learn to make decisions and outline patterns from data – dates back at least to the 1950s. However, a recent shift towards data-driven approaches and the advent of deep learning methods have led to significant advances over the past two decades.[1] Deep learning is a subset of machine learning that involves the extraction of many feature layers from raw data and that utilizes neural networks, which have been likened to the animal nervous system to produce complex predictive outputs (**Figure 1**).[2] In medicine, deep learning has been applied to a diverse array of clinical problems, from the detection of diabetic retinopathy, to the detection of breast cancer on standard mammogram to the diagnosis of cutaneous malignancy.[3]

The field of gastroenterology has been an early leader in bridging the gap between artificial intelligence model development and clinical trial validation. Machine learning and deep learning have been applied in many realms of gastroenterology. In endoscopy, it has been used anywhere from optical biopsy and polyp detection during colonoscopy,[4, 5] to the diagnosis of H. pylori and gastric cancer during upper endoscopy,[6, 7] to the automatic detection and classification of lesions during video capsule endoscopy (VCE).[8-10] One of the first randomized trials utilizing artificial intelligence in clinical medicine was in gastroenterology and entailed the application of a deep-learning-based computer aided detection (CADe) algorithm for the automatic detection of polyps during colonoscopy.[11] AI efforts outside of gastrointestinal endoscopy have focused on predictive modeling in inflammatory bowel disease (IBD),[1] irritable bowel syndrome (IBS),[12] and pancreaticobiliary disease for both diagnosis and to augment therapeutic management.[13, 14] In addition, gastroenterologists, important stakeholders in the conversation and potential end-users of these AI tools, have a strong interest and generally positive attitude towards AI applications in gastroenterology according to early surveys in the United States.[15]

As AI applications for clinical medicine advance rapidly, there is a clear need for guidance surrounding AI-specific study design, evaluation, comparison, analysis and reporting of results. Several initiatives are in the publication or pre-publication phase including AI-specific

amendments to minimum reporting guidelines for clinical trials, society task force initiatives aimed at priority use cases and research priorities and minimum reporting guidelines that guide the reporting of clinical prediction models. In this paper we will first examine applications of AI in clinical trials within gastroenterology and discuss the elements of newly published checklists that are intended to inform the design and reporting of future clinical trials. We will then review AI-applications at the pre-trial level in both endoscopy and other subfields of gastroenterology and explore areas where further guidance is needed to supplement the current guidance available at the pre-trial level.

THE CURRENT STATE OF CLINICAL AI TRIALS IN GASTROENTEROLOGY

Within gastroenterology, most prospective work has focused on computer aided detection (CADe) and computer-aided diagnosis (CADx) during colonoscopy. CADe involves the automatic detection of polyps during colonoscopy and CADx, or optical biopsy, involves the prediction of polyp histology without the need for tissue biopsy. For both CADx and CADe, early efforts in the 1990s involved traditional machine-learning techniques with explicit feature extraction methods, with algorithms trained and validated on still images captured from colonoscopy video.[3] The introduction of deep learning led to significant improvements in algorithm performance in both subfields.[5, 16, 17] Early studies involving deep learning for CADe and CADx involved the publication of training and validation data for a given algorithm on still images, then retrospective video and finally on prospective video. In 2019, Wang et al. published the first randomized trial utilizing artificial intelligence in clinical medicine. In this study, 1058 patients in a single center in China were randomized to receive diagnostic colonoscopy with or without the assistance of a CADe system on a second monitor. Investigators found a significant increase in adenoma detection rate (ADR), 20.3% in the control arm and 29.1% in the experimental arm ($p<0.001$), as well as an increase in the mean number of adenomas.[11] Similar studies in China, including a double blind randomized clinical trial have found similar increases in ADR.[18, 19] The same authors also published a randomized tandem colonoscopy trial and found a lower adenoma miss rate in AI-assisted colonoscopy compared to high definition white light colonoscopy.[20]

Repici et. al published the first multi-center randomized controlled trial examining a similar AI intervention to previous authors in China (a deep learning algorithm projected on a second screen intended to aid the endoscopist in the detection of polyps). This study also showed a significant increase in ADR (54.8% vs. 40.4 %) with a relative risk [RR] of 1.30 (95% CI, 1.14-1.45) in a provider-participant population with a higher baseline ADR and in a more homogenous patient population presenting for screening or surveillance colonoscopy. Authors found no significant increase in withdrawal time between groups and no significant increase in resection of non-significant lesions.[21] In a meta-analysis of 5 of these randomized trials, Barua et al found an ADR of 29.6% (95% CI 22.2-37.0) for AI-assisted colonoscopy versus 19.3% (95% CI 12.7-25.9) for colonoscopy without AI.[22] In line with these positive results in AI for colonoscopy, a number of CADe and CADx systems for colonoscopy have cleared regulatory approval in certain regions of the world and are starting to be distributed on the market (e.g. GI-Genius, Medtronic; CAD-EYE, Fujifilm; DISCOVERY, Pentax; EndoBRAIN-EYE, Olympus; ENDO-AID, Olympus; WISE VISION, NEC Corporation).[23]

EXPERT GUIDANCE ON REPORTING OF AI-SPECIFIC CLINICALTRIAL DESIGN

The majority of these studies were published before any guidance surrounding AI-specific trial design and reporting of outcomes was available. One of the first guidelines designed for implementation at the trial level are AI-specific extensions to the Consolidated Standards of Reporting Trials (CONSORT) and Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT) checklists. The original CONSORT and SPIRIT statements are widely used evidence-based recommendations for the reporting of randomized controlled trials (CONSORT) and the development of trial protocols (SPIRIT).[24, 25] In 2020, Liu and Cruz Rivera et al. published AI-specific extensions developed using a Delphi methodology with an international multi-disciplinary consortium of AI experts.[26-28] They include AI-specific items such as explicit statement of the intended role of the AI intervention, description of the AI-human interaction and explicit reporting of inclusion and exclusion criteria at the level of input data as well as at the level of the participant. While these are minimum reporting guidelines, they represent an important step forward for the field, and while they are generally applicable to all trials examining an AI intervention, they also fit well within the canon of current and expected

work in GI endoscopy. **Table 1** details best practices for AI research in gastroenterology and includes examples drawn from the CONSORT-AI and SPIRIT-AI statements as well as from a variety of other sources.

In part because of the rapid progress examining CADe and CADx technologies in GI endoscopy, major societies are also starting to put forth priority statements and suggested guidelines for AI research. In recent guidelines for advanced imaging in the detection and differentiation of colorectal neoplasia, the European Society of Gastrointestinal Endoscopy (ESGE) suggest the possible incorporation of CADe and CADx technologies in colonoscopy.[29] In 2020, the American Society for Gastrointestinal Endoscopy (ASGE) published a position statement on priorities for AI progress in gastrointestinal endoscopy [30]. This includes anticipated needs for computer vision in GI endoscopy, decision support, practice management, data storage and prospective validation.[30]

THE CURRENT STATE OF MACHINE LEARNING AT THE PRE-TRIAL LEVEL

The field of gastroenterology has taken an early role in clinical trial efforts for AI with the publication of multiple randomized trials in the last two years. However, the majority of published work over the past decade consists of retrospective and prospective studies at the pre-trial level.

Applications in Endoscopy and Imaging

Computer vision has been applied successfully to a wide range of endoscopic modalities from video capsule endoscopy to endoscopic ultrasound. One of the early applications of deep learning in GI endoscopy was in CADx or optical biopsy. Recent prospective work has shown the potential to accurately differentiate between adenomatous and non-adenomatous polyp histology in-situ and potentially avoid the need for biopsy or resection of diminutive polyps in the rectosigmoid colon.[4, 31]

While CADe and CADx systems have been studied most extensively in colonoscopy, we are starting to see the application of similar technologies to upper endoscopy as well. In a meta-analysis of 23 studies, Lui et al. found relatively high areas under the receiver operating characteristic curve (AuROC) for the detection of stomach neoplasia, Barrett's esophagus,

squamous esophagus and H. pylori, though this work is early and the analysis was based on retrospective studies using still images.[32] In a recent meta-analysis of 19 studies related to upper GI neoplasia, Arribas et al. similarly found encouraging test characteristics for the detection of squamous cell neoplasia. Barrett's esophagus-related and gastric adenocarcinoma, but found overall low study quality with a high risk of selection bias.[33] Deep learning has also been used to successfully classify pathology into adenocarcinoma, adenoma and non-neoplastic for upper-GI biopsies;[6] celiac disease versus environmental enteropathy versus normal;[34] and in automating endoscopic severity scores in ulcerative colitis.[35]

In video capsule endoscopy (VCE), we are also starting to see the application of CADe and CADx algorithms. Deep learning algorithms have been applied successfully for the detection of protruding lesions in the small bowel,[8] inflammation, ulcers, polyps, parasites,[9] and celiac disease.[36] Deep learning has also been applied to the field of therapeutic endoscopy, such as in the detection and characterization of focal liver lesions on endoscopic ultrasound,[37] the differentiation between autoimmune pancreatitis and pancreatic cancer,[14] and the characterization of pancreatic cyst fluid.[38] Artificial intelligence methods have also been successfully utilized in endoscopy training and quality assurance from the analysis of bowel prep adequacy[39] to the reduction of blind spots during upper endoscopy[40] to optimizing the quality of colonoscopy.[41] In medical imaging specific to gastroenterology, early applications include the automatic segmentation of CT enterography images in Crohn's disease in order to predict stricturing versus non-stricturing disease.[42]

Beyond Endoscopy: Other Applications of AI in Gastroenterology

While many recent advances have been in computer vision as applied to medical imaging and technology, investigators have also begun to successfully apply machine learning to a variety of clinical questions within gastroenterology. One area of emerging success is in applying machine learning to precision medicine in IBD. Machine learning has been used to successfully analyze sources of big data from the electronic health record to imaging to high throughput omics data in order to tease out patterns and make predictions in IBD.[1] Waljee et al. developed a predictive model using 20,368 Veterans Health Administration based on a random forest (RF) algorithm to predict a combined endpoint of outpatient corticosteroid use and hospitalizations as a surrogate for IBD flare. Authors found a high AuROC of 0.87 and found several important

predictors including previous hospitalization and corticosteroid use.[43] Random forests have also been used to differentiate fecal bacteria in active vs. remission states.[44] In addition, significant recent progress has been made developing models used to predict and evaluate endoscopic severity in Crohn's disease and ulcerative colitis. Bossuyt et al. developed a novel endoscopic severity score using a computer algorithm (red density) used to predict endoscopic and histologic severity. The resultant RD algotihm correlated with Roberts histological index, Mayo endoscopic subscore and UC Endoscopic severity index.[45] Takenaka et. al developed a deep neural network trained on 40,758 colonoscopy images and 6885 biopsy results from patients with a confirmed diagnosis of ulcerative colitis. They then tested the resultant deep learning algorithm prospectively on 875 patients with UC. The system identified patients in endoscopic remission defined as a UC Endoscopic Index of Severity (UCEIS) score of 0 with an accuracy of 90.1% (95% [CI] 89.2%-90.9%) when compared to expert endoscopist analysis as the gold standard. The system also accurately predicted histologic remission.[35] In a follow-up study, authors showed that the same algorithm could predict patient prognosis in relation to UC-related hospitalization and need for colectomy favorably when compared to human experts.[46] Other, similar systems have been developed to assess endoscopic severity in UC. [47, 48]Early efforts in other arenas have been aimed toward the discovery of new therapies, the identification of disease sub-groups, the prediction of drug response and the improvement of diagnosis.[1]

Outside of the world of IBD, machine learning techniques are starting to be applied for predictive modeling in other disease states. Tap et al. collected fecal and mucosal samples from patients who met criteria for IBS and used a machine learning procedure to generate a microbial signature for severe versus mild IBS patients.[12] Jovanovic et. al examined 291 consecutive patients who presented to the hospital with suspected choledocholithiasis. They developed a conventional multivariate regression model and an artificial neural network and compared each model's performance for the prediction of positive findings on resultant ERCP. They found an AuROC of 0.884 for the neural network versus an AuROC of 0.787 for the multivariate logistic regression prediction model.[13] Kudo et al. developed a prediction model based on an artificial neural network which used 8 pre-operative variables to predict the presence of lymph node metastasis in T1 colorectal cancer. The constructed model outperformed current U.S. guidelines in identifying patients with T1 colorectal cancers who had lymph node metastases.[49] Recently,

Shung et al. developed a machine learning model that outperformed existing clinical risk scoring systems for determining risk in patients presenting with upper GI bleed.[50]

EXPERT GUIDANCE AND FUTURE DIRECTIONS AT THE PRE-TRIAL LEVEL

While standardized, thoughtful design and transparent reporting at the level of prospective randomized clinical trials is the ultimate goal, as we can see from the numerous examples mentioned above, the majority of current publications examining AI in gastroenterology are at the pre-trial level. It is equally important that initial development and validation studies as well as studies examining resultant technologies in both retrospective and non-randomized prospective settings are conducted and reported with standardized guidance as well. Currently, however; there is little guidance in this area. The Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD) statement was published in 2015 and offers guidance surrounding key reporting items in the development, evaluation and improvement of conventional prediction models.[51] It includes items specific to conventional prediction models such as presentation of the full prediction model with regression coefficients and intercept, report of model performance and discussion of potential clinical use but also includes general items that may be extrapolated to models based on machine learning and deep learning techniques (**Table 1**).

Despite the potential for extrapolation, few current studies applying artificial intelligence to clinical medicine utilize these best practices. In a systematic review examining design, reporting standards, risk of bias and study claim versus reality, Nagendran et al. analyzed 81 non-randomized studies comparing a deep learning algorithm in medical imaging with clinician performance. Authors used a modified version of the TRIPOD statement to generally assess adherence to reporting standards and also applied the prediction model risk of bias assessment tool (PROBAST) to assess for the risk of bias.[52] They found that adherence to reporting standards was poor and overall publications adhered to 24%-90% of TRIPOD items with a median of 62% (interquartile range of 45-69%). In addition, they found a high risk of bias in 72% of non-randomized studies.[52] At the time of this writing, there is an initiative to develop an extension of the TRIPOD guidelines specific to machine learning, the TRIPOD-ML statement.[53]

Our hope is that this will encourage researchers to develop and report on ML-based prediction models and other AI-based technologies in a standardized, transparent fashion.

Other groups are also working on best practices and suggestions for more transparent, reproducible, ethical and effective (TREE) ML research. Vollmer et al., for example, outlined 20 key questions that are intended to be a framework for researchers and readers of AI research and are also intended to be a checklist for editors and peer reviewers to use as a starting point for the evaluation of the quality of a given manuscript.[54] Essential questions such as those generated by this group are essential to all stakeholders in AI research from developers to clinical researchers to journal editors and peer-reviewers and should be examined critically before the implementation of AI algorithms in clinical practice (**Table 2**).[55]

CONCLUSION

We are at a time of exciting opportunity in AI research in gastroenterology, with the recent publication of multiple high quality, randomized trials examining the role of computer vision in GI endoscopy. However, there are concerns that early successes and media popularization of deep learning may lead to the rapid implementation of AI in clinical medicine without thoughtful, standardized and transparent algorithm development and reporting. Recent guidance from the CONSORT and SPIRIT steering groups in the form of AI-specific extensions to previous statements are a monumental step forward, but this is not enough. Design and reporting at the pre-trial level must be examined and standardized as well. In addition, the methods with which AI algorithms are developed and compared must be critically examined before implementation is considered ethical or feasible. For example, we need standardization of the study and terminology of CADe and CADx algorithms in clinical use, we need publicly available data for the development of new algorithms, and we need methods to directly compare emerging systems. We are at a time of unprecedented growth and excitement for the potential that artificial intelligence and deep learning may unlock in the field of gastroenterology. Indeed, there is little doubt that AI has the potential to impact nearly every aspect of clinical gastroenterology, and meaningful progress will require a responsible and systematic approach towards research investigation.

Figures and Tables Legend

Figure 1. Overview of definitions

Table 1. Best practices in artificial intelligence research and examples in the literature

Table 2. Some barriers to implementation of AI in clinical practice, consequences and potential solutions

Table 1. Best practices in artificial intelligence research and examples in the literature

Best Practices	Examples in the literature
Title. Indicate that the intervention involves artificial intelligence/machine learning and specify the type of model; Specify the intended use of the AI intervention.†‡	"Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study"[19]
Background and objectives. Specify the objectives, including whether the study describes the development or validation of the model, or both.°	"We aimed to develop an AI-assisted polyp detection system and to validate its performance using a large colonoscopy video database designed to be publicly accessible."[56]
Outcome. Clearly define the outcome that is predicted by the prediction model, including how and when assessed.°	"The primary outcome was a composite measure capturing both use of outpatient corticosteroids prescribed for IBD and inpatient hospitalizations associated with a diagnosis of IBD."[43]
Eligibility Criteria. State the inclusion and exclusion criteria at the level of participants.†‡	"The target population included 40- to 80-year-old subjects undergoing colonoscopy for primary CRC screening or post-polypectomy surveillance, as well as for workup following fecal immunohistochemical test (FIT) positivity...or for symptoms/signs. Patients were excluded in case of personal history of CRC, or inflammatory bowel disease, previous colonic resection, antithrombotic therapy precluding polyp resection, and lack of informed written consent"[21]

Eligibility Criteria. State the inclusion and exclusion criteria at the level of the input data.†‡	"From the recorded subjects, we excluded 1.) those diagnosed with inflammatory bowel disease, 2.) those diagnosed with polyposis disease, 3.) nonepithelial lesions, 4.) polyps recorded on only low-quality frames with artifact, and 5.) lesions not recorded with white-light endoscopy" [4 *]
Interventions. Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements.‡	"We fully integrated CADe in the endoscopy system, completely mimicking the usual routine of the operators by overimposing the CADe box over the same endoscopic screen." [21] "The system was connected to the endoscopy generator, and the video stream was captured synchronously. Furthermore, the system processed each frame and displayed the detected polyp location with a hollow blue tracing box on an adjacent monitor with a simultaneous sound alarm (figure 1) (see online supplementary file 1). The system was turned on during withdrawal only." [11]
Interventions. Specify whether there was human–AI interaction in the handling of the input data, and what level of expertise was required of users.‡	"Eight physicians from the division of gastroenterology participated in the study, including two senior endoscopists (>20000 colonoscopies), two midlevel endoscopists (between 3000 and 10000 colonoscopies) and four junior endoscopists (between 100 and 500 colonoscopies)… The system was turned on during withdrawal only. The endoscopist focused mainly on the main monitor during the procedure and was prompted to look at the system monitor by the sound alarm. The endoscopist was required to check every polyp location detected by the system." [11]

Missing data. Describe how missing data were handled (for example, complete-case analysis, single imputation, multiple imputation) with details of any imputation method.°	"Missing lab covariate values were imputed based on the median value of the lab from all the previous visits. Patients missing more than 50% of lab data were excluded from analysis."[43]
Development versus validation. For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.°	"In total, 56,668 images were used for the machine-learning. These training frames were divided into 2 categories, training images and validation images, which aimed to tune multiple parameters of YoloV3. In the study, 51 899 frames were used as training images and the remaining 4769 images as validation images."[56]

† SPIRIT-AI Checklist [28]

‡ CONSORT-AI Checklist [26]

° Elements from the TRIPOD statement generalizable to AI research

Table 2. Some barriers to implementation of AI in clinical practice, consequences and potential solutions

Possible Barriers to Implementation	**Consequence**	**Potential Solutions**
Heterogeneity in quality of data used for model training and validation – E.g. Missing data, irrelevant data	Overfitting of a given model on training/validation data leads to decreased performance in the real-world setting	Minimize missing data, ensure robust validation on internal and external sources of data that are

		separated in time and space, ensure that the ground truth for the development of a given algorithm is generalizable
Lack of ability to directly compare models from different research groups	Parallel development and publication of multiple models based on similar or differing technologies from a number of groups with no means of differentiating each model	1. Explicit statement of training and validation procedures 2. Making data and model publicly available [54] 3. Head-to-head comparison of models in randomized clinical trials (may not be practical) 4. Transparent reporting of performance statistics (e.g. sensitivity, specificity, positive predictive value, misclassification, ROC) 5. Standardization of clinical definitions (e.g false positive definition in the study of CADe) 6. Development of high-quality data sets designed to serve as a

		benchmark for comparison of multiple models [56]
Inappropriate comparisons between a given algorithm to a clinical baseline	"Weak comparator bias" [54] wherein the benefits of an AI algorithm is overstated as a result of comparison to sub-par competitors (e.g. overstatement of improvement in ADR by comparing a CADe system to novice endoscopists in colonoscopy)	Compare the model to the relevant clinical gold standard
Low uptake and/or low engagement for a given algorithm despite proven benefit	Underutilization of potentially useful technology	1. Involvement of multiple stakeholders for a given technology including patients, developers, commercial entities, physicians, physician societies, regulatory bodies and policymakers 2. Focus on availability, accessibility, cost and personalization
Ambiguity surrounding liability in cases where AI may cause harm	Confusion around fault in cases where harm is attributed to artificial intelligence-based systems	Adaptation of product liability law to fit the landscape of AI in clinical medicine
Potential exacerbation of inequities in gender, sex and ethnicity	A given algorithm may make disproportionate errors in different populations [54]	Include key populations in development data to increase predictive accuracy within subgroups

References

1. Seyed Tabib NS, Madgwick M, Sudhakar P, Verstockt B, Korcsmaros T, Vermeire S. Big data in IBD: big progress for clinical practice. *Gut*. 2020; **69**: 1520-32.
2. Chartrand G, Cheng PM, Vorontsov E*, et al.* Deep Learning: A Primer for Radiologists. *Radiographics*. 2017; **37**: 2113-31.
3. Alagappan M, Brown JRG, Mori Y, Berzin TM. Artificial intelligence in gastrointestinal endoscopy: The future is almost here. *World J Gastrointest Endosc*. 2018; **10**: 239-49.
4. Mori Y, Kudo SE, Misawa M*, et al.* Real-Time Use of Artificial Intelligence in Identification of Diminutive Polyps During Colonoscopy: A Prospective Study. *Ann Intern Med*. 2018; **169**: 357-66.
5. Misawa M, Kudo SE, Mori Y*, et al.* Characterization of Colorectal Lesions Using a Computer-Aided Diagnostic System for Narrow-Band Imaging Endocytoscopy. *Gastroenterology*. 2016; **150**: 1531-2 e3.
6. Iizuka O, Kanavati F, Kato K, Rambeau M, Arihiro K, Tsuneki M. Deep Learning Models for Histopathological Classification of Gastric and Colonic Epithelial Tumours. *Sci Rep*. 2020; **10**: 1504.
7. Itoh T, Kawahira H, Nakashima H, Yata N. Deep learning analyzes Helicobacter pylori infection by upper gastrointestinal endoscopy images. *Endosc Int Open*. 2018; **6**: E139-E44.
8. Saito H, Aoki T, Aoyama K*, et al.* Automatic detection and classification of protruding lesions in wireless capsule endoscopy images based on a deep convolutional neural network. *Gastrointest Endosc*. 2020; **92**: 144-51 e1.
9. 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**: 1044-54 e5.

10. Glissen Brown JR, Berzin TM. Deploying artificial intelligence to find the needle in the haystack: deep learning for video capsule endoscopy. *Gastrointest Endosc*. 2020; **92**: 152-3.
11. Wang P, Berzin TM, Glissen Brown JR*, et al.* Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut*. 2019; **68**: 1813-9.
12. Tap J, Derrien M, Törnblom H*, et al.* Identification of an Intestinal Microbiota Signature Associated With Severity of Irritable Bowel Syndrome. *Gastroenterology*. 2017; **152**: 111-23.e8.
13. Jovanovic P, Salkic NN, Zerem E. Artificial neural network predicts the need for therapeutic ERCP in patients with suspected choledocholithiasis. *Gastrointest Endosc*. 2014; **80**: 260-8.
14. Marya NB, Powers PD, Chari ST*, et al.* Utilisation of artificial intelligence for the development of an EUS-convolutional neural network model trained to enhance the diagnosis of autoimmune pancreatitis. *Gut*. 2020.
15. Wadhwa V, Alagappan M, Gonzalez A, Chandnani M, Berzin TM. 542 Gastroenterologist Sentiment Toward Artificial Intelligence (AI) in Endoscopic Practice: A Nationwide Survey. *Official journal of the American College of Gastroenterology | ACG*. 2019; **114**: S313.
16. Wang P, Xiao X, Glissen Brown JR*, et al.* Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nat Biomed Eng*. 2018; **2**: 741-8.
17. Urban G, Tripathi P, Alkayali T*, et al.* Deep Learning Localizes and Identifies Polyps in Real Time With 96% Accuracy in Screening Colonoscopy. *Gastroenterology*. 2018; **155**: 1069-78 e8.
18. Gong D, Wu L, Zhang J*, et al.* Detection of colorectal adenomas with a real-time computer-aided system (ENDOANGEL): a randomised controlled study. *Lancet Gastroenterol Hepatol*. 2020; **5**: 352-61.
19. Wang P, Liu X, Berzin TM*, et al.* Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. *Lancet Gastroenterol Hepatol*. 2020; **5**: 343-51.

20. Wang P, Liu P, Glissen Brown JR*, et al.* Lower Adenoma Miss Rate of Computer-aided Detection-Assisted Colonoscopy vs Routine White-Light Colonoscopy in a Prospective Tandem Study. *Gastroenterology*. 2020.

21. Repici A, Badalamenti M, Maselli R*, et al.* Efficacy of Real-Time Computer-Aided Detection of Colorectal Neoplasia in a Randomized Trial. *Gastroenterology*. 2020.

22. Barua I, Vinsard D, Jodal H*, et al.* Artificial Intelligence for Polyp Detection during Colonoscopy: A Systematic Review and Meta-Analysis. *Endoscopy*. 2020.

23. Yamada M, Saito Y, Imaoka H*, et al.* Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy. *Sci Rep*. 2019; **9**: 14465.

24. Begg C, Cho M, Eastwood S*, et al.* Improving the quality of reporting of randomized controlled trials. The CONSORT statement. *JAMA*. 1996; **276**: 637-9.

25. Chan AW, Tetzlaff JM, Altman DG*, et al.* SPIRIT 2013 Statement: defining standard protocol items for clinical trials. *Rev Panam Salud Publica*. 2015; **38**: 506-14.

26. Liu X, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI Extension. *BMJ*. 2020; **370**: m3164.

27. Bilal M, Brown JRG, Berzin TM. Incorporating standardised reporting guidelines in clinical trials of artificial intelligence in gastrointestinal endoscopy. *The Lancet Gastroenterology & Hepatology*.

28. Cruz Rivera S, Liu X, Chan AW*, et al.* Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. 2020; **26**: 1351-63.

29. Bisschops R, East JE, Hassan C*, et al.* Advanced imaging for detection and differentiation of colorectal neoplasia: European Society of Gastrointestinal Endoscopy (ESGE) Guideline - Update 2019. *Endoscopy*. 2019; **51**: 1155-79.

30. Berzin TM, Parasa S, Wallace MB, Gross SA, Repici A, Sharma P. Position statement on priorities for artificial intelligence in GI endoscopy: a report by the ASGE Task Force. *Gastrointest Endosc*. 2020.

31. Horiuchi H, Tamai N, Kamba S, Inomata H, Ohya TR, Sumiyama K. Real-time computer-aided diagnosis of diminutive rectosigmoid polyps using an auto-fluorescence imaging system and novel color intensity analysis software. *Scand J Gastroenterol*. 2019; **54**: 800-5.

32. Lui TKL, Tsui VWM, Leung WK. Accuracy of artificial intelligence-assisted detection of upper GI lesions: a systematic review and meta-analysis. *Gastrointest Endosc*. 2020.

33. Arribas J, Antonelli G, Frazzoni L*, et al.* Standalone performance of artificial intelligence for upper GI neoplasia: a meta-analysis. *Gut*. 2020.

34. 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.

35. 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**: 2150-7.

36. Zhou T, Han G, Li BN*, et al.* Quantitative analysis of patients with celiac disease by video capsule endoscopy: A deep learning method. *Comput Biol Med*. 2017; **85**: 1-6.

37. Marya NB, Powers PD, Fujii-Lau L*, et al.* Application of artificial intelligence using a novel eus-based convolutional neural network model to identify and distinguish benign from malignant hepatic masses. *Gastrointest Endosc*. 2020.

38. Kurita Y, Kuwahara T, Hara K*, et al.* Diagnostic ability of artificial intelligence using deep learning analysis of cyst fluid in differentiating malignant from benign pancreatic cystic lesions. *Sci Rep*. 2019; **9**: 6893.

39. Zhou J, Wu L, Wan X*, et al.* A novel artificial intelligence system for the assessment of bowel preparation (with video). *Gastrointest Endosc*. 2020; **91**: 428-35 e2.

40. Wu L, Zhang J, Zhou W*, et al.* Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. *Gut*. 2019; **68**: 2161-9.

41. Thakkar S, Carleton NM, Rao B, Syed A. Use of Artificial Intelligence-Based Analytics From Live Colonoscopies to Optimize the Quality of the Colonoscopy Examination in Real Time: Proof of Concept. *Gastroenterology*. 2020; **158**: 1219-21 e2.

42. 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*. 2020; **26**: 734-42.

43. Waljee AK, Lipson R, Wiitala WL*, et al.* Predicting Hospitalization and Outpatient Corticosteroid Use in Inflammatory Bowel Disease Patients Using Machine Learning. *Inflamm Bowel Dis*. 2017; **24**: 45-53.

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

45. Bossuyt P, Nakase H, Vermeire S*, et al.* Automatic, computer-aided determination of endoscopic and histological inflammation in patients with mild to moderate ulcerative colitis based on red density. *Gut*. 2020; **69**: 1778-86.

46. Takenaka K, Ohtsuka K, Fujii T, Oshima S, Okamoto R, Watanabe M. Deep neural network accurately predicts prognosis of ulcerative colitis using endoscopic images. *Gastroenterology*. 2021.

47. 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**: 416-21 e1.

48. 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.

49. Kudo S-e, Ichimasa K, Villard B*, et al.* Artificial Intelligence System to Determine Risk of T1 Colorectal Cancer Metastasis to Lymph Node. *Gastroenterology*.

50. Shung DL, Au B, Taylor RA*, et al.* Validation of a Machine Learning Model That Outperforms Clinical Risk Scoring Systems for Upper Gastrointestinal Bleeding. *Gastroenterology*. 2020; **158**: 160-7.

51. Collins GS, Reitsma JB, Altman DG, Moons KG. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ*. 2015; **350**: g7594.

52. Nagendran M, Chen Y, Lovejoy CA*, et al.* Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ*. 2020; **368**: m689.

53. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet*. 2019; **393**: 1577-9.

54. Vollmer S, Mateen BA, Bohner G*, et al.* Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *BMJ*. 2020; **368**: l6927.

55. van der Sommen F, de Groof J, Struyvenberg M*, et al.* Machine learning in GI endoscopy: practical guidance in how to interpret a novel field. *Gut*. 2020; **69**: 2035-45.

56. Misawa M, Kudo SE, Mori Y*, et al.* Development of a computer-aided detection system for colonoscopy and a publicly accessible large colonoscopy video database (with video). *Gastrointest Endosc*. 2020.

Table 1. Best practices in artificial intelligence research and examples in the literature

† SPIRIT-AI Checklist [28]

Best Practices	Examples in the literature
Title. Indicate that the intervention involves artificial intelligence/machine learning and specify the type of model; Specify the intended use of the AI intervention.†‡	"Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study"[19]
Background and objectives. Specify the objectives, including whether the study describes the development or validation of the model, or both.°	"We aimed to develop an AI-assisted polyp detection system and to validate its performance using a large colonoscopy video database designed to be publicly accessible."[56]
Outcome. Clearly define the outcome that is predicted by the prediction model, including how and when assessed.°	"The primary outcome was a composite measure capturing both use of outpatient corticosteroids prescribed for IBD and inpatient hospitalizations associated with a diagnosis of IBD."[43]
Eligibility Criteria. State the inclusion and exclusion criteria at the level of participants.†‡	"The target population included 40- to 80-year-old subjects undergoing colonoscopy for primary CRC screening or post-polypectomy surveillance, as well as for workup following fecal immunohistochemical test (FIT) positivity...or for symptoms/signs. Patients were excluded in case of personal history of CRC, or inflammatory bowel disease, previous colonic resection, antithrombotic therapy precluding polyp resection, and lack of informed written consent"[21]

Eligibility Criteria. State the inclusion and exclusion criteria at the level of the input data.†‡	"From the recorded subjects, we excluded 1.) those diagnosed with inflammatory bowel disease, 2.) those diagnosed with polyposis disease, 3.) nonepithelial lesions, 4.) polyps recorded on only low-quality frames with artifact, and 5.) lesions not recorded with white-light endoscopy" [4 *]
Interventions. Describe how the AI intervention was integrated into the trial setting, including any onsite or offsite requirements.‡	"We fully integrated CADe in the endoscopy system, completely mimicking the usual routine of the operators by overimposing the CADe box over the same endoscopic screen." [21] "The system was connected to the endoscopy generator, and the video stream was captured synchronously. Furthermore, the system processed each frame and displayed the detected polyp location with a hollow blue tracing box on an adjacent monitor with a simultaneous sound alarm (figure 1) (see online supplementary file 1). The system was turned on during withdrawal only." [11]
Interventions. Specify whether there was human–AI interaction in the handling of the input data, and what level of expertise was required of users.‡	"Eight physicians from the division of gastroenterology participated in the study, including two senior endoscopists (>20000 colonoscopies), two midlevel endoscopists (between 3000 and 10000 colonoscopies) and four junior endoscopists (between 100 and 500 colonoscopies)… The system was turned on during withdrawal only. The endoscopist focused mainly on the main monitor during the procedure and

‡ CONSORT-AI Checklist [26]

	was prompted to look at the system monitor by the sound alarm. The endoscopist was required to check every polyp location detected by the system." [11]
Missing data. Describe how missing data were handled (for example, complete-case analysis, single imputation, multiple imputation) with details of any imputation method.°	"Missing lab covariate values were imputed based on the median value of the lab from all the previous visits. Patients missing more than 50% of lab data were excluded from analysis."[43]
Development versus validation. For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.°	"In total, 56,668 images were used for the machine-learning. These training frames were divided into 2 categories, training images and validation images, which aimed to tune multiple parameters of YoloV3. In the study, 51 899 frames were used as training images and the remaining 4769 images as validation images."[56]

° Elements from the TRIPOD statement generalizable to AI research

Table 2. Some barriers to implementation of AI in clinical practice, consequences and potential solutions

Possible Barriers to Implementation	**Consequence**	**Potential Solutions**
Heterogeneity in quality of data used for model training and validation – E.g. Missing data, irrelevant data	Overfitting of a given model on training/validation data leads to decreased performance in the real-world setting	Minimize missing data, ensure robust validation on internal and external sources

		of data that are separated in time and space, ensure that the ground truth for the development of a given algorithm is generalizable
Lack of ability to directly compare models from different research groups	Parallel development and publication of multiple models based on similar or differing technologies from a number of groups with no means of differentiating each model	1. Explicit statement of training and validation procedures 2. Making data and model publicly available [54] 3. Head-to-head comparison of models in randomized clinical trials (may not be practical) 4. Transparent reporting of performance statistics (e.g. sensitivity, specificity, positive predictive value, misclassification, ROC) 5. Standardization of clinical definitions

		(e.g false positive definition in the study of CADe) 6. Development of high-quality data sets designed to serve as a benchmark for comparison of multiple models [56]
Inappropriate comparisons between a given algorithm to a clinical baseline	"Weak comparator bias" [54] wherein the benefits of an AI algorithm is overstated as a result of comparison to sub-par competitors (e.g. overstatement of improvement in ADR by comparing a CADe system to novice endoscopists in colonoscopy)	Compare the model to the relevant clinical gold standard
Low uptake and/or low engagement for a given algorithm despite proven benefit	Underutilization of potentially useful technology	1. Involvement of multiple stakeholders for a given technology including patients, developers, commercial entities, physicians, physician societies, regulatory bodies and policymakers 2. Focus on availability, accessibility, cost and personalization

Ambiguity surrounding liability in cases where AI may cause harm	Confusion around fault in cases where harm is attributed to artificial intelligence-based systems	Adaptation of product liability law to fit the landscape of AI in clinical medicine
Potential exacerbation of inequities in gender, sex and ethnicity	A given algorithm may make disproportionate errors in different populations [54]	Include key populations in development data to increase predictive accuracy within subgroups

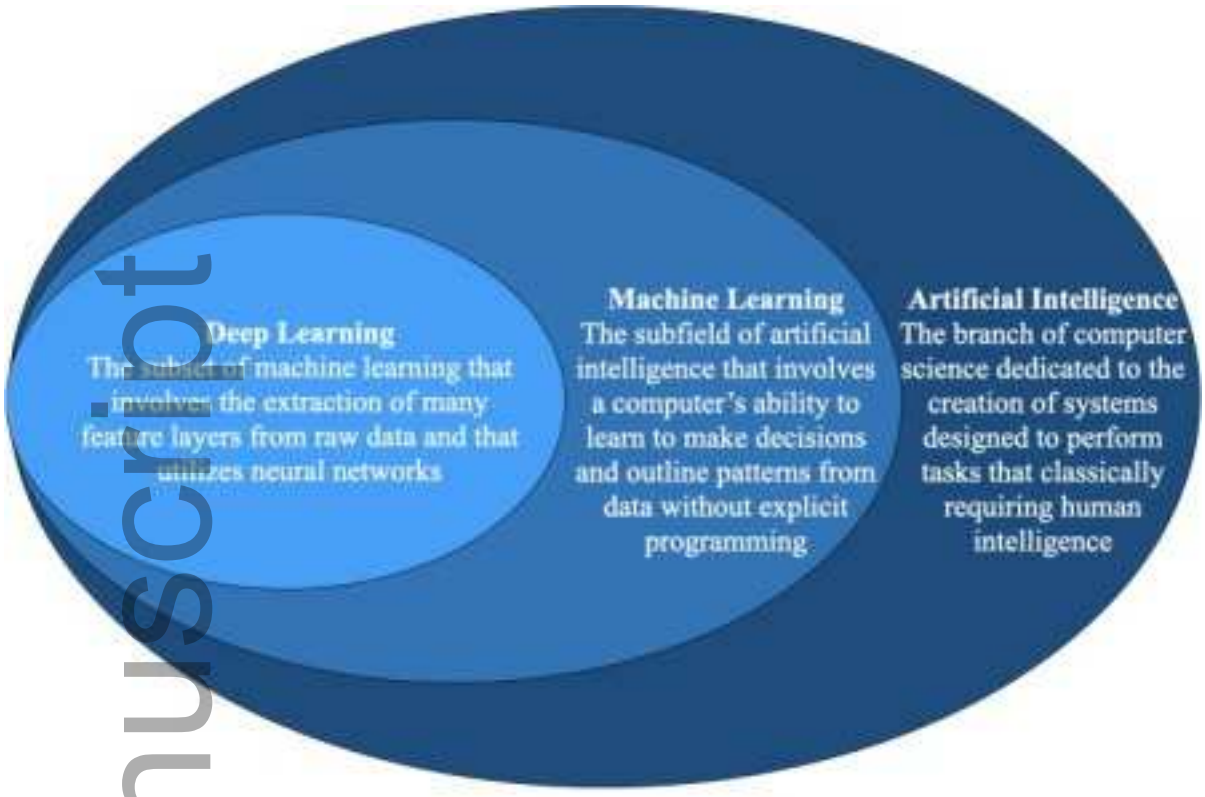

Figure 1. Overview of definitions

den_13974_f1.tiff