Endoscopy and central reading in inflammatory bowel disease clinical trials: achievements, challenges and future developments

Klaus Gottlieb ,[1] Marco Daperno,[2] Keith Usiskin,[3] Bruce E Sands,[4] Harris Ahmad ,[5] Colin W Howden,[6] William Karnes,[7] Young S Oh,[8] Irene Modesto,[9] Colleen Marano,[10] Ryan William Stidham ,[11] Walter Reinisch [12]

► Additional material is published online only. To view please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2020-320690).

For numbered affiliations see end of article.

Correspondence to
Dr Klaus Gottlieb, Immunology, Eli Lilly and Company, Indianapolis, Indiana 46225, USA; klaus.gottlieb@lilly.com

Received 15 January 2020
Revised 4 June 2020
Accepted 13 June 2020
Published Online First
22 July 2020

To cite: Gottlieb K, Daperno M, Usiskin K, *et al*. *Gut* 2021;**70**:418–426.

ABSTRACT

Central reading, that is, independent, off-site, blinded review or reading of imaging endpoints, has been identified as a crucial component in the conduct and analysis of inflammatory bowel disease clinical trials. Central reading is the final step in a workflow that has many parts, all of which can be improved. Furthermore, the best reading algorithm and the most intensive central reader training cannot make up for deficiencies in the acquisition stage (clinical trial endoscopy) or improve on the limitations of the underlying score (outcome instrument). In this review, academic and industry experts review scoring systems, and propose a theoretical framework for central reading that predicts when improvements in statistical power, affecting trial size and chances of success, can be expected: Multireader models can be conceptualised as statistical or non-statistical (social). Important organisational and operational factors, such as training and retraining of readers, optimal bowel preparation for colonoscopy, video quality, optimal or at least acceptable read duration times and other quality control matters, are addressed as well. The theory and practice of central reading and the conduct of endoscopy in clinical trials are interdisciplinary topics that should be of interest to many, regulators, clinical trial experts, gastroenterology societies and those in the academic community who endeavour to develop new scoring systems using traditional and machine learning approaches.

INTRODUCTION

Central reading, that is, independent, off-site, blinded review or reading of imaging endpoints in clinical trials, used in other disease areas for decades, came to inflammatory bowel disease (IBD) clinical trials only recently. It was first reported in a meeting abstract in 2006[1] and gained traction in 2013 after Feagan *et al*[2] showed the importance of central reading in IBD clinical trials. Central reading is the final step in a workflow that has many parts, all of which can be improved more easily than scoring systems (outcome instruments).

Central reading has generally been successful by promoting objectivity, lowering variability, reducing the placebo response rate and increasing the effect size of active drug but challenges remain. For example, the higher effect sizes for centrally read studies has been questioned recently.[3] Placebo remission rates are lower with central reading but there is considerable variability between studies, affecting point estimates for sample size calculations or the increasing screen failure rates.

Central reading can be implemented in different ways. We propose a framework that predicts when we can expect to see improvements in statistical power, affecting trial size and chances of success. Artificial intelligence methods are expected to make important contributions to accuracy, precision and reproducibility of central reading. Some of the connected issues, for example, how to train computational scoring systems, will be addressed in this paper.

The endoscopic scoring systems (outcome instruments) are at the centre of clinical trial endoscopy and they require improvements, but these will take years. Still, problems that arise at the instrument level are difficult to mitigate with central reading and we will put our thoughts about better scoring systems in context with our other recommendations.

In addition, seemingly mundane but important organisational and operational factors, such as training and retraining of readers, optimal bowel preparation for colonoscopy, video quality, optimal or at least acceptable read duration times, and other quality control subjects need to be addressed as well.[4]

ACQUISITION STAGE

Standardisation of the bowel prep and choice of colonoscopy versus sigmoidoscopy in ulcerative colitis

Clinical trial protocols leave the bowel prep, and, in case of ulcerative colitis (UC), also the choice of instrument, up to the discretion of the principal investigator (PI). This may be the reason suboptimal videos due to poor bowel prep or insufficient washing are a significant problem in clinical trials. Suboptimal videos may lead to missing data, if they are considered unreadable, false interpretations by the central reader or an increased chance for discrepant reads if more than one reader is part of the read algorithm. Diligent washing of the mucosa by the endoscopist is also necessary when the bowel prep is otherwise good in order to wash of fibrin exudates which could otherwise either masquerade as ulcers or obscure them.

The administration of the first half of the preparation the evening before colonoscopy and the second half in the morning of the procedure (so called split-dose regimen) has shown superior efficacy in bowel cleansing over the original regimen of administering the whole preparation the day before the procedure and as such has become part of guidelines.[5] In practice, more than half of patients do not take the second half of the prep when they are scheduled for the procedure before 10:00 a.m. The fear of incontinence on the way to the endoscopy service and the refusal to wake up in the very early morning to complete the bowel preparation represent the main barriers against split dosing.[6] The practical consequence is that they may have a suboptimal prep, but better education may help.[7]

Head-to-head studies of bowel preps have only recently become available. Gu *et al* found in a large non-commercial 'real-world' prospective multicenter trial with 4339 colonoscopies and 75 endoscopists that MiraLAX with Gatorade, MoviPrep and Suprep were associated with superior tolerability and bowel cleansing.[8] As tolerability of bowel preps should be optimised for clinical trial participants with active IBD, bowel preps should be selected accordingly. Polyethylene glycol 3350 and some form of an electrolyte balanced sports drink may be ubiquitously available, and this combination may also be optimal because PEG based bowel prep regimens seem to have the lowest rate of bowel prep induced mucosal artefacts.[9]

A related issue is the choice of procedure in UC clinical trials, colonoscopy or sigmoidoscopy. Kato *et al* performed a retrospective analysis using data collected at a university hospital and demonstrated that up to 27% of patients with UC colonoscopy showed more severe lesions situated in the descending colon compared with the sigmoid or rectum.[10] Divergent results were reported in a post hoc analysis of endoscopic examinations from the EUCALYPTUS trial of etrolizumab in UC.[11] The use of sigmoidoscopy only to confirm mucosal healing was associated with a risk of underestimating disease activity and overestimating treatment efficacy when endoscopic healing was defined as an an endoscopic Mayo Score (eMS) of 0 or 1, but not if it was defined as eMS=0.

If sigmoidoscopy is chosen, an enema prep may come along with it. While there are few data, enema preps may be inferior to colonoscopy preps. Some believe that sigmoidoscopy is more acceptable to patients, neglecting that it is mostly performed without sedation. Colonoscopy, in contrast, is a procedure almost universally performed with moderate to deep sedation. Indeed, limited data seem to show that patients find sigmoidoscopy more difficult.[12] In addition, a sedated procedure may allow a more thorough examination. We believe that if not colonoscopy, then a colonoscopy bowel prep should be the standard for clinical trials.

The site endoscopist is responsible for video quality

It is universally agreed that videos submitted by site endoscopists are of variable quality and there are multiple different reasons for this. High-definition white light endoscopy was introduced in 1993 and is the current standard in gastrointestinal (GI) endoscopic practice and has replaced standard-definition video endoscopy and is required for participation in IBD clinical trials because the image quality is demonstrably better.[13] However, we are concerned that several image vendors do not actually make high-definition videos available to the central readers, instead videos are downsampled to mediocre resolutions of 640 × 424 pixels for reasons that are unclear.

Another factor is that the length of the videos varies considerably even if colonoscopies and sigmoidoscopies are considered separately. It is well known that the time spent inspecting the colonic mucosa is correlated with the likelihood of finding or missing polyps, and longer withdrawal times are associated with a reduced incidence of interval cancer after screening colonoscopy.[14] The measurement of colonoscopy withdrawal time has therefore become one of the indispensable quality indicators for colonoscopy[15] and some such metric adopted for IBD clinical trials could be used both for the site endoscopist and for the central reader in reviewing a video.

Site endoscopists, that is, investigators who personally perform the colonoscopy, control the quality of the video and the biopsies at the source by ensuring the best possible bowel preparation, diligent washing during the procedure and appropriate insufflation, attention to withdrawal time, keeping an adequate distance from the mucosa, the recording of relevant lesions, obtaining biopsies according to protocol. In the past, they have also supplied the endoscopic score.

It was then suggested that PIs may be too biassed to be scorers,[2] and their role has been diminished to that of a videographer for the central reader(s). It has been previously argued that bias can be diminished or abolished if site endoscopists know that their score will be compared with those of one or more central readers.[16] If well trained in the scoring algorithm, site endoscopists could also be an integral part of the reading algorithm. The benefits of using site endoscopist as readers in one trial has been reported by Reinisch *et al*.[17]

Trained site readers could stay current by acting as central readers for other clinical trials. Development and standardisation of needed training programmes, perhaps delivered by electronic means and open to all who are qualified, not only on the scoring system but also on withdrawal technique, time, washing, could best be organised by the GI societies. Increasing screen failure rates in IBD clinical trials[18] could perhaps in part be mitigated by more fully engaging the site endoscopist/PI.

Training of site personnel

Central reading service vendors typically take charge of the training of ancillary site personnel as the procedures for video capture and electronic transfer differ from vendor to vendor. Clinical research associates provided by other vendors or the sponsor are often a conduit for training and trouble shooting. Ancillary personnel could be better used, for example, by assisting site endoscopists in the proper recording of biopsy locations, biopsy protocol adherence, adherence to recommended insertion and withdrawal techniques and duration, even an understanding of the scoring system could improve team performance. This training could be delivered electronically, just in time, or baseline with refresher just before a scheduled visit, and may also be sponsor or vendor agnostic. Endoscopy technicians could become Clinical Trial Endoscopy Specialists, akin to the emerging role of the Clinical Trial Imaging Specialist and supervise training and performance of several clinical trial sites.

HISTORICAL AND PRACTICAL CONSIDERATIONS ABOUT ENDOSCOPIC ACTIVITY SCORING SYSTEMS

The multitude of endoscopic scoring systems in IBD has periodically been reviewed.[19–22] They are typically called endoscopic disease severity scores or indices, but it is not certain what exactly it is they measure (in this section).

Attempts of endoscopic scoring in UC have a longer tradition than in Crohn's disease and as such endoscopic disease assessment measures are a regular, integral part of combined endpoints in clinical trials on UC since the 1980s. Currently, both the European Medicines Agency and the US Food and Drug Administration (FDA) endorse the eMS[23] as endoscopic assessment tools for drug development in UC, although it appears that for the former also the Ulcerative Colitis Endoscopic Index of Severity (UCEIS) is acceptable. Neither of the agencies put great emphasis on the development of a new endoscopic assessment system for UC.

The eMS is similar to the Baron score[24] established for the use of the 25 cm Lloyd-Davies rigid sigmoidoscope to record phenomena of increasing bowel inflammation, erythema, erosions and ulcers in patients with UC.[25] Their assessment was limited by the insertion depth of the instrument, patient tolerance and field of view. More than 50 years later, descendants and modifications of this score have survived. Currently, the eMS is the dominant endoscopic scoring instrument in UC, likely owing to broad physician familiarity and ease of use. The eMS, proposed in 1987 for clinical trials in UC, seeks to categorise UC endoscopic activity using a 0–3 categorical scale of disease severity based on the presence and gestalt of the endoscopic features, erythema, vascular pattern, friability, erosion, ulcer and spontaneous bleeding.[25] More recently, FDA has prompted a modification of the eMS in a way that a value of 1, which is the endoscopic endpoint criterion for endoscopic improvement, formerly mucosal healing, does no longer include friability, but only erythema and abnormal vascular pattern.[23]

The eMS has never been subjected to a proper validation process and, historically, the scoring system was primarily aimed at highlighting responsiveness to drugs, specifically, 5-aminosalicylic acid compounds. The eMS intrinsically lacks the ability to precisely depict the spectrum of endoscopic severity. It remains to be determined whether the endoscopic features determining the high ranges, that is, friability, erosions, ulcers and spontaneous bleeding, are independent signatures of incremental endoscopic severity of UC or expression of the phenotypic heterogeneity of disease, as encountered by central readers receiving cases from across the globe. Furthermore, the discrimination of the features spontaneous bleeding and friability by a central reader necessitates the full recording of endoscopy, which is not always available. Spontaneous bleeding can be solely observed during the advancement phase of endoscopy, whereas friability is assessed by the presence of patches of superficial blood caused by trauma from the endoscopic procedure and only observed during withdrawal. In addition to the uncertainty of independence of the features of eMS, the lack of dynamic range and the challenges of separating neighbouring severity grades resulting in limited interobserver agreement have been criticised (see table 1).

Even though the eMS has not been developed as a prognostic tool and face validity of the endoscopic features defining the lower range of the score (0–1), erythema and/or abnormal vascular pattern only, is elusive, endoscopic improvement commonly defined as eMS ≤1, and complete endoscopic healing (ie, eMS=0) are associated with superior disease outcomes, including avoidance of colectomy.[26] In addition, due to its wide use and its categorical score, easy algorithms for central reading, discussed below, have been established, although data on the impact of various reader paradigms on point estimates of placebo remission/response rates and effect sizes are limited.[17]

In an attempt to address some of the limitations of the eMS, the UCEIS was developed in the late 2000s and is the product of a validation process.[27 28] The UCEIS individually grades vascular pattern, erosions and ulceration, and bleeding, resulting in an expanded range (0–8) and a more pronounced sensitivity to change with endoscopic remission defined as a UCEIS of 0. It has been shown to have more reproducibility compared with eMS, although inter-reader and intrareader agreement of the rectal bleeding component is limited. Nevertheless, the UCEIS is based on components subsumed by the eMS and therefore, is still essentially based on subjective feature classification, for which independent pathogenetic relevance is unclear.[24] A strong correlation between the UCEIS and the eMS has been shown,[29] however, the UCEIS appears to more accurately identify severe cases strengthening the UCEIS prognostic value for improved long-term outcome for a UCEIS ≤1.[30]. So far, the use of the UCEIS in clinical trials is limited and reader paradigms on defining agreement and adjudication are more complex for this more granular score as compared with the 4-category eMS.

A factor in most UC scoring systems is that the total extent of disease at baseline or on follow-up may not be known. This is not the case in Crohn's disease in which endoscopic scoring systems stipulate a colonoscopy. For example, the eMS and UCEIS both score the worst endoscopic lesion without consideration of the extent of the mucosa involved. In clinical trials, a full colonoscopy is often performed at baseline followed by a sigmoidoscopy at the efficacy visits. The disease may have receded, however, this improvement will not be captured in the scores if an ulcer is still be present in the rectosigmoid. Because most UC endoscopic scoring approaches do not account for the mucosal inflammatory load in UC, this might explain in part suboptimal correlations between endoscopic disease severity and levels of objective biomarkers. A worst-lesion scoring system could mask clinically relevant endoscopic responses. In central reading, there are scenarios where the mucosal surface affected by the most severe lesion is impressively diminished between visits but without triggering a change in the overall score if only a small area of signifying severe lesion is still left. For example, in a recent trial of induction therapy with etrasimod in UC the Spearman correlation coefficient of fecal calprotectin with the eMS was 0.32 for placebo, 0.29 for the 1 mg dose and 0.70 for the 2 mg dose[31]

There have been attempts to adopt segmental scoring similarly to Crohn's disease. The Modified Mayo Endoscopic Score is calculated on the basis of the eMS in five colonic segments[32] and the Degree of Ulcerative colitis Burden of Luminal INflammation score combine extent and severity of the disease according to the eMS.[33] The development and validation of a scoring system that is a better proxy of the inflammatory burden in UC by documenting the total extent of endoscopic abnormalities would require complete colonoscopy. Such a score could be expected to require fewer patients to show a response to an intervention, and in combination with histology, could be a significant advance in the development of UC outcome instruments.

The Crohn's Disease Endoscopic Index of Severity (CDEIS)[34] and its simplified counterpart, the Simple Endoscopic Score for Crohn's Disease (SES-CD)[35] were developed and validated in 1989 and in 2004, respectively, with the intent to offer a numeric transformation of a precise severity reporting. The CDEIS is based on four domains: deep ulcerations (weighted by a factor of 12), superficial ulcerations (weighted by a factor of 6, surface involved by disease (assessed by cm Visual Analogue Scale), and surface involved by ulcerations (assessed by cm Visual Analogue Scale). Each domain is measured in five ileocolonic segments: the rectum, sigmoid and left colon, transverse colon, right colon and ileum. To the sum score of the individual segments divided by the number of assessed segments, the presence of stenosis, either as a result of ulcer or not, is added. The possible scores are ranging from 0 to 44.

Table 1 Comparison of strengths and limitations of commonly used endoscopic scores

Score		Endoscopic activity reporting	Responsiveness to treatments	Prognostic value	Central reading
eMS	**Pros**	▶ Gross classification of the gestalt of inflammation. ▶ Present standard for Drug Agencies (FDA, EMA).	▶ Development focused at responsiveness. ▶ Extensively used over past 20 years in trials.	▶ Limited data for a prognostic role in the literature.	▶ Algorithms for central reading. ▶ Categorical score leads to easier algorithms for adjudication. ▶ Widely used over past 5 years.
	Cons	▶ Final score defined by worst lesion. ▶ Lacks precision for global burden of severity and extent of lesions. ▶ Lack of face validity ▶ Endoscopic features only post hoc defined. ▶ Limited spectrum at lower and higher spectrum of activity.	▶ Lack of ability to highlight segmental healing. ▶ Lack of responsiveness due to limited range.	▶ Not developed with prognostic intent.	▶ Limited interobserver agreement. ▶ Inconsistencies between readers if insufficient washing of the mucosa. ▶ Data on impact of reader paradigms on eMS-based endpoints is missing.
UCEIS	**Pros**	▶ Extensive characterisation and validation of elemental endoscopic lesions focused at agreement.	▶ Better range than eMS.		▶ Already used in some trials.
	Cons	▶ Lacks precision for global burden of severity and extent of lesions.	▶ Lack of ability to highlight segmental healing. ▶ Limited use in clinical trials. ▶ Development not focused at responsiveness.	▶ Not developed with prognostic intent.	▶ Agreement and adjudication more complex for more granular scores as compared with categorical scores. ▶ Modest agreement on some lesions (eg, bleeding).
Rutgeerts	**Pros**	▶ Clear-cut description of elemental lesions.		▶ Development focused on prognosis. ▶ Prognostic value has been reproduced.	▶ Central reading easy to implement Algorithms for eMS easily exportable to Rutgeerts' score.
	Cons	▶ Not an activity measure. ▶ Does not evaluate endoscopic activity outside of the anastomotic site.	▶ No responsiveness evaluation.	▶ Developed for end-to-end anastomoses, never validated for side-to-side anastomoses. ▶ Limited interobserver agreement.	▶ Limited interobserver agreement. ▶ No data on impact of read paradigm on outcome.
CDEIS	**Pros**	▶ Developed and validated in order to precisely report disease activity.	▶ Shown in few trials, even if not explicitly developed for responsiveness.		▶ Used in clinical trials. ▶ Excellent inter-rater reliability.
	Cons	▶ Complexity. ▶ Exact weight of each variable to be better clarified. ▶ Unvalidated thresholds for remission and response. ▶ The definition of remission does not exclude the presence of ulcers.	▶ Not developed with focus on responsiveness.	▶ Limited prognostic value of the sum score.	▶ Agreement and adjudication more complex for continuous scores as compared with categorical scores. ▶ Not developed for postoperative anatomy.
SES-CD	**Pros**	▶ Developed and validated in order to precisely report disease activity. ▶ Possibility to easily exclude a given variable. ▶ Segmental and ulcer subscores can be calculated.	▶ Shown in several trials, even if not explicitly developed for responsiveness.		▶ Widely used in trials. ▶ Excellent inter-rater variability. ▶ Different reader algorithms available (fix or sliding scale for adjudication, paired reading …).
	Cons	▶ Relative complex. ▶ Exact weight of each variable to be better clarified. ▶ Unvalidated thresholds for remission and response.	▶ Not developed with focus on responsiveness.	▶ Limited prognostic value of sum score.	▶ Agreement and adjudication more complex for more granular scores as compared with categorical scores. ▶ No adjustment for missing segments due to sum score. ▶ Not developed for postoperative anatomy.

CDEIS, Crohn's Disease Endoscopic Index of Severity; EMA, European Medicines Agency; eMS, endoscopic Mayo Score; FDA, Food and Drug Administration; SES-CD, Simple Endoscopic Score for Crohn's Disease; UCEIS, Ulcerative Colitis Endoscopic Index of Severity.

The SES-CD score is based on the same five ileocolonic segments, but accounts for the size of mucosal ulcers (0–3), the ulcerated surface (0–3), the affected surface (0–3) and the presence of passable or non-passable stenosis (0–3). In contrast to the CDEIS, the SES-CD is a simple sum score for the assessed segments and possible scores are ranging from 0 to 56. For both scores higher numbers are indicative of greater degrees of mucosal disease activity. The clinical adoption of these measures has been slow as a result of the calculation requirements, but despite validated in fewer studies than CDEIS, SES-CD, which correlates with CDEIS, is easier to use and the primary endoscopic disease severity tool endorsed by regulatory agencies.

Inter-reader agreement of both, CDEIS and SES-CD, is excellent,[36] however, the lack of validated thresholds for the definitions of endoscopic remission and response remains a major issue. For CDEIS endoscopic remission is arbitrarily defined by a score <3, a cut-off that does not exclude the presence of ulcers and is, therefore not in line with the aspired treatment target of absence of ulcers in Crohn's disease.[37] Consequently, the International Organisation of IBD committee review on clinical trials defined endoscopic remission either as lack of ulcerations or SES-CD ≤2, the latter also precluding ulcers.[38] Similarly, endoscopic response is also based on arbitrarily chosen thresholds of a ≥50% decrease in SES-CD or CDEIS which may, however, correlate with corticosteroid free remission.[39]

CDEIS and SES-CD intrinsically lack prognostic implications, and they were not originally developed for evaluation of responsiveness (even if they were subsequently shown to be quite reliable also for analysis of pretreatment/post-treatment responsiveness with a slight benefit of the SES-CD over the CDEIS[40]); the SES-CD may present the additional advantage to allow for easy evaluation of segmental and ulceration subscores separately from the total score, while for CDEIS this is not possible (see table 1). None of the scores has been developed to adjust for the postoperative anatomy in Crohn's disease and its associated specific lesions, for example, at the anastomotic ring as well as changes in segmental transition zones. The impact of read paradigms on SES-CD defined endpoints have been studied and are discussed below.

The Rutgeerts score was developed with prognostic intent in 1990 for postoperative Crohn's disease recurrence,[41] and leaving apart the issue of lacking a formal validation, it was not intended to describe endoscopic severity precisely, and intrinsically lacks any precision with respect to responsiveness.

While the literature is variable, it has mostly shown that many endoscopic scores do not correlate well with patient-related outcomes[42] or even other 'objective' disease markers such as faecal calprotectin. Conceptually, there are many reasons for this lack of correlation, residing either with the dependent or independent variable or both. Perhaps basing scores on abdominal pain and stool frequency is too reductionist, perhaps patients are remiss in recording their symptoms properly, perhaps psychological comorbidities interfere.[43] Importantly, what we think is more or less objective and reliable, the endoscopic scores—developed for endoscopic instruments that have long become obsolete—may potentially not quite measure what is relevant, as described above. For example, we do not know to what extent the currently used endoscopic scores reflect the underlying biology. Preliminary data show that in UC histological indices track gene expression changes by several orders of magnitude better than the UCEIS or eMS.[27]

Concepts of mucosal improvement and healing could perhaps be approached more holistically, that is by integrating histological, endoscopic, and transcriptomics perspectives.

New endoscopy scores, and possibly, new histological scores, could be developed using machine learning (ML) that will help us surmount our cognitive limitations. Unsupervised learning (ie, not conditioned on human reader scores) could uncover endoscopic features that have either escaped our attention or are too difficult to evaluate, creating score that are more granular approaching a continuous scale. Clearly, ML as applied to colonoscopic disease activity scores has potential, but for supervised learning, the likely point of departure, a reliable reference standard for algorithm development is needed.

THE CENTRAL READ PROCESS

Baseline central reader qualifications

In April 2019,[18] there were 48 384 patients and 13 762 sites participating in IBD clinical trials (data from ICON Clinical Research Organisation). In consequence there is a large demand for central readers and competition for qualified candidates is increasing.

Qualifications of physicians who treat patients with IBD and who perform endoscopy vary, and for imaging core labs that serve the regulatory needs of pharmaceutical companies, qualifications must go beyond some basic items.[4] Currently it is required that central readers have an up-to-date curriculum vitae, are board certified in gastroenterology and document a variable number of years of postcertification experience in treating IBD patients. We are not aware that there is a requirement for a minimum number of colonoscopies during the most recent year in practice, as for example required for recredentialing at the Mayo Clinic in Rochester, Minnesota,[44] and elsewhere in the USA. It remains to be determined whether the adenoma detection rate,[45] as a valuable proxy for the endoscopist's effort, diligence and commitment to quality (compulsiveness) in performing screening colonoscopies,[46] could also be helpful in selecting appropriate candidates to read clinical trial colonoscopies in IBD.

Central reading vendors enrol candidate central readers in proprietary reader training programmes. Typically, they include the assessment of training video cases according to the independent review charter which is proposed by the vendor and approved by the clinical trial sponsor. Intrareader and inter-reader metrics are used to identify outliers for retraining. Nonetheless, once a central reader is qualified, periodic retraining is necessary as scoring behaviour might shift over time. The selection criteria for central readers, the actual implementation of the central read according to a number of different approaches—to be discussed below—and the variability of training are all factors that should be further explored, and standardised, a task the GI societies maybe best equipped to handle.

Training and qualification of readers on the scoring system

It has been shown that interobserver agreement among experienced physicians who are only instructed but not trained in the scoring system can be quite poor,[47] but, fortunately, training can improve these rates significantly. Daperno *et al*[48] used a templated training programme that consisted of slide and video clip presentation with experienced IBD faculty as instructors. The attendees were all gastroenterologists or internists with a minimum postcertification experience of 3 years and a maximum experience of 30 years, and all were actively involved both in IBD clinics and in endoscopy, similar to the qualifications needed for a central reader pool. The inter-rater agreement increased from kappa 0.51 (95% CI 0.48 to 0.55) to 0.76 (95% CI 0.72 to 0.79) for the Mayo endoscopic subscore, and from 0.45 (95% CI 0.40 to 0.50) to 0.79 (0.74 to 0.83) for the Rutgeerts score before and after the training programme, respectively, and both differences were significant (p<0.0001).

Central reading companies have their own proprietary training programmes which are often briefly and in very general terms described in the independent review charter. These charters summarise the vendor's interpretation of a published score, which usually leaves too much latitude for implementation, contributing to intrareader and inter-reader disagreement. However, impressive inter-reader and intrareader agreement from training programmes may not necessarily reflect bona fide inter-reader convergence but might instead be heavily influenced by the quality of videos and the magnitude of ambiguity of the mock cases to be assessed. A tendency to avoid the difficult has, for example, been reported for peer review in radiology, 'where easy cases were often chosen'.[49] Therefore, those metrics might not necessarily reflect the performance of readers in the actual study situation where video quality could be suboptimal to poor and the interpretation of borderline cases more contentious.

Central reading and ML

The application of ML methods to image analysis, frequently termed computer vision, can offer opportunities to replicate expert endoscopic scoring standard with high reproducibility, accuracy and precision. Automated systems could be trained using libraries of digital endoscopic videos collected in the course of clinical trials, paired with their respective centrally reviewed endoscopic scores. Early efforts attempting to replicate expert scoring have shown promise, though interpretation of unaltered endoscopic video demands more training than disease severity alone.[50] Negotiating variable bowel preparations, disambiguating spontaneous versus procedure induced tissue changes (eg, bleeding), managing variations in the non-standardised video recording, digital compression, and addressing difference in endoscopist withdrawal patterns are all performed intuitively be experienced human reviewers but still present challenges for machines.

While there may be ongoing controversy about the best central read algorithm, those that use statistical data aggregation (see next section) seem to be best suited for ML development. A possible approach would consist of forming a precompetitive consortium where pharmaceutical manufacturers supply the videos to be reread according to uniform criteria by qualified readers organised through GI societies, who are also active in sponsoring reader training programmes. Additionally, computational methods for standardised central scoring could also provide more informative quantitative statistics on the confidence in the predicted endoscopic score, thereby quantifying the ambiguity that still occurs even between trained reviewers. Finally, perfect replication of endoscopic scoring by computational methods will also perfectly replicate the biases and error of scoring used for training. A perfect training set does not and will not exist; careful thought to minimising bias and understanding the error of the ground truth selected for training will be essential.

CENTRAL READING ALGORITHMS: STATISTICAL VERSUS NON-STATISTICAL

Read algorithms are different from the endoscopic scoring system. They formalise how, exactly, given a specific scoring system, readers (scorers, evaluators) should conduct the reading/image evaluation, and how the final scoring results for a given instance is to be arrived at, especially when there is more than one reader assigned per read instance, which is current practice in late stage trials.

A more detailed discussion of many practically important aspects of central read algorithms can be found in the (online supplementary appendix).

In brief, given the substantial inter-reader disagreement, attempts have been made to somehow combine the assessment of more than one well-trained reader for a final score. For this type of data aggregation different methods can be used. In principle, the methods can be divided into statistical data aggregation techniques, which by mathematical necessity result in improved accuracy compared with one central reader models, and non-statistical (social) data aggregation methods, where accuracy gains cannot be predicted, because interpersonal dynamics do not necessarily result in improved accuracy.

In a consensus-based approach, a panel looks at the image or other matter of interest and comes, after open deliberation, to a conclusion. This process cannot be described mathematically as the inclination or power of individuals to influence others can neither be predicted nor easily measured. How a consensus process for central reading can be counterproductive when applied to IBD clinical trials, has previously been illustrated.[51]

Another non-statistical approach is that of adjudication. When two people cannot agree, they ask a third person to be an adjudicator. If used correctly, the word adjudication means that the third person knows the assessment of the other parties and takes it under consideration, the decision is final with the 'judge'. This is in distinction to an anonymous process where there is equal weighting of each reader's score, that is, voting. The same as above applies, the dynamics of this process cannot be described mathematically.

In contrast, averaging and voting are statistical data aggregation methods. Here, accuracy improvements are transparent. For averaging, they follow a square root law which holds that the SE of the sample mean decreases with the square root of the number of samples.[52] In contrast, scores which have only few levels, such as the eMS (0,1,2,3) cannot be properly averaged, because the distances between ordinal numbers are unknown.[53] Still, statistical data aggregation can be done using voting. The accuracy improvements using voting can also be described mathematically with the Condorcet Jury Theorem.[16] Voting algorithms use two readers, and, in cases of disagreement, an optional third reader (2+1 reader algorithm). In case reader 1 and 2 agree, the score is final. If not, reader number 3 votes, independently, in other words, without knowing that there was a disagreement. Reader 3 is not an adjudicator, but another voter, see Gottlieb and Hussain[16] and Ahmad *et al*.[54]

QUALITY CONTROL AND ONGOING MONITORING OF CENTRAL READER COMPETENCY

A review of typical reading charters reveals that retraining and retesting is envisioned, often on an annual basis, but learning theory would suggest that refreshers should be done when needed and should coincide with new reading tasks or sessions. Reader quality is often assessed by evaluating reader performance using interobserver statistics. Whether the statistic chosen is Cohen's kappa or the intraclass correlation coefficient (ICC) does not matter, as both metrics are trying to condense multiple levels of information into a single statistic which can be problematic if context is not kept in mind. For example, kappa (and ICC) values change with the prevalence of disease, and, as was recently shown in simulation, spurious kappa changes can occur during different phases of a clinical trial, even if the actual reader performance is kept constant.[55] Practically speaking, kappa metrics before and after an intervention may not be comparable, and they should only be compared during the same phase of the study. Such statistics may also differ between active and placebo without representing changes in reader performance.

Another area which has escaped attention scrutiny is the influence image quality has on interobserver agreement. It makes sense to postulate that as image quality declines, mostly because of a suboptimal bowel preparation and inadequate washing by the colonoscopist, observer agreement should decline as well. While there are, to our knowledge, no comparable studies in IBD central reading, this effect has been described in other imaging fields.[49] So far, little or no attention is placed by central reading vendors on assessing image quality or bowel prep quality on a reproducible basis and no thresholds have been defined when

Table 2 Summary of suggested changes and improvements in the conduct of clinical trial endoscopy

Suggested improvement or change (in order of presentation in the paper)	Importance	Ease of implementation
Colonoscopy only for UC trials.	++	++
Require split dosing for colonoscopy preps.	+++	+++
Avoid early morning colonoscopy for trial participants.	++	++
Standardise bowel prep to polyethylene glycol 3350.	+++	+++
Require vendors to present videos to central readers at the same resolution as recorded (no downsampling).	+++	+++
Capture metrics for colonoscopy acquisition times (site reader) and viewing times (central reader) and set minimum standards.	++	++
Involve site endoscopists as readers.	++	+
Central reading training programmes by GI societies.	+++	+
Better training and collaborative use of ancillary personnel.	+++	+
Design new scoring systems (endoscopic outcome instruments), especially for UC, that better reflect inflammatory burden and are validated for their context of use, possibly using machine learning.	++++	+
Harmonise central reader qualification processes with clinical credentialing requirements.	++	++
Insist on more transparency regarding vendor central reader training programmes and harmonisation (see also above 'Central reading training programmes by GI societies').	+++	++
Embrace ML to inform development of new scoring systems.	+++	+
Read algorithms (aggregation of the input of more than one reader per video into the final score): choose statistical over non-statistical data aggregation methods.	++++	+++
Create prespecified thresholds for acceptable versus unacceptable bowel preps, possible implementation with ML algorithms prior to presentation to central readers.	+++	++

GI, gastrointestinal; ML, machine learning; UC, ulcerative colitis.

a video is objectively unreadable. There are now AI algorithms available that can score bowel prep quality without reader input and they could be adopted to central reading work flows.[56] This is attractive because the algorithm could be deployed early on during the acquisition of the colonoscopy video, allowing the PI to institute immediate remedial action, for example, re-prep the patient for the following day.

CONCLUSIONS

There are many important components that can make clinical trial endoscopy and central reading more accurate. The best reading-algorithm and the most intensive central reader training cannot make up for deficiencies in the acquisition stage or improve on the limitations of the underlying score. Here we have discussed multiple areas of possible improvement, some of which can be implemented quickly or easily, others which will require further research and extensive development (table 2).

We believe that the one-central reader model is problematic and that that multi-reader models can best be conceptualised along the lines of whether they are statistical or non-statistical (social). Only the former promises reproducible performance gains. The statistical fundamentals of central reading seem to be clear, but there remain many questions at the margins that need to be resolved.

One way forward is for Pharmaceutical companies to make deidentified and annotated (scored) videos available for training purposes and ML projects, and GI societies could serve as the independent intermediaries. ML will eventually alleviate many of the issues now encountered in central reading, that is, time commitment, reader variability and bias, motivation, and fatigue, and will allow better scoring systems to be developed. In addition, withdrawal time, prep quality and inflammation, integrated over the entire withdrawal phase of colonoscopy, could be quantified algorithmically.

Central reading is too important for the future development of GI therapeutics, especially in IBD, to be left to proprietary approaches. Instead, industry, academia and GI societies need to take concerted action in propelling the science forward and help establish reader training programmes.

Author affiliations
[1]Immunology, Eli Lilly and Company, Indianapolis, Indiana, USA
[2]A.O. Ordine Mauriziano di Torino, Torino, Italy
[3]Immunology, Celgene Corp, Summit, New Jersey, USA
[4]Dr Henry D Janowitz Division of Gastroenterology, Mount Sinai School of Medicine, New York, New York, USA
[5]Immunoscience, Bristol-Myers Squibb Co, New York, New York, USA
[6]Gastroenterology, Univ Tennessee, Memphis, Tennessee, USA
[7]Gastroenterology, UC Irvine, Irvine, California, USA
[8]Immunology, Genentech Inc, South San Francisco, California, USA
[9]Inflammation & Immunology, Pfizer Inc, New York, New York, USA
[10]Janssen Research & Development, Spring House, Pennsylvania, USA
[11]Internal Medicine, University of Michigan, Ann Arbor, Michigan, USA
[12]Department of Medicine IV, Medical University Vienna, Vienna, Austria

Contributors KG, in conjunction with WR, devised the project, the main conceptual ideas and wrote the first and subsequent drafts and circulated them among all coauthors. The coauthors revised the drafts in an iterative fashion until consensus was achieved. All authors participated in the drafting and all approved the manuscript and secured organisational permission where required.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests Klaus Gottlieb is an employee and stockholder of Eli Lilly and Company. Marco Daperno reports personal fee from AbbVie, Chiesi, Ferring, Janssen, MD, Pfizer, and Quintiles, grants and personal fees from Takeda, and nonfinancial support from SOFAR and Mundifarma.Keith Usiskin is an employee of Celgene now Bristol Myers Squibb. Bruce Sands has received personal fees from AbbVie, Akros Pharma, Amgen, Arena Pharmaceuticals, AstraZeneca, Boehringer-Ingelheim, Forward Pharma, Bristol-Myers Squibb, Immune Pharmaceuticals, Shire, Synergy Pharmaceuticals, Theravance Biopharma R&D, TiGenix, TopiVert Pharma, Receptos, Allergan, EnGene, Target PharmaSolutions, Lycera, Lyndra, Ironwood Pharmaceuticals, Salix; grants, personal fees and non-financial support from Celgene, Takeda, Pfizer, Janssen, personal fees and non-financial support from Prometheus Laboratories, Hoffman-La Roche, MedImmune, Lilly, Vivelix Pharmaceuticals, UCB, Oppilan Pharmaceuticals, Gilead, Rheos Medicines, Seres Therapeutics, 4D Pharma, Capella Bioscience, Otsuka, Ferring, Protagonist Therapeutics, Palatin Technologies. Harris Ahmad is an employee of Bristol-Myers Squibb. Colin Howden reports being a consultant for Phathom Pharmaceuticals, Ironwood, RedHill Biopharma, Alfasigma, OtsukaStockholder in Antibe Therapeutics andCo-Editor of Alimentary Pharmacology & Therapeutics.William Karnes reports that he is CMO, cofounder, shareholder and paid contractor for Docbot.Young Oh is an employee of Genentech, a member of the Roche Group and a Roche stock holder. Irene Modesto is and employee and stockholders of Pfizer. Colleen Marano is an employee of Janssen. Ryan Stidham reports consultancy for Abbvie, Janssen, Merck, Takeda. Investigator initiated research support from Abbvie. University of Michigan has filed a provisional patent on behalf of RWS related to machine learning for endoscopic evaluation in IBD.Walter Reinisch has served as a speaker for Abbott Laboratories, Abbvie, Aesca, Aptalis, Astellas, Centocor, Celltrion, Danone Austria, Elan, Falk Pharma, Ferring, Immundiagnostik, Mitsubishi Tanabe Pharma Corporation, MD, Otsuka, PDL, Pharmacosmos, PLS Education, Schering-Plough, Shire, Takeda, Therakos, Vifor, Yakult, as a consultant for Abbott Laboratories, Abbvie, Aesca, Algernon, Amgen, AM Pharma, AMT, AOP Orphan, Arena Pharmaceuticals, Astellas, Astra Zeneca, Avaxia, Roland Berger, Bioclinica, Biogen IDEC, Boehringer-Ingelheim, Bristol-Myers Squibb, Cellerix, Chemocentryx, Celgene, Centocor, Celltrion, Covance, Danone Austria, DSM, Elan, Eli Lilly, Ernest & Young, Falk Pharma, Ferring, Galapagos, Genentech, Gilead, Grünenthal, ICON, Index Pharma, Inova, Janssen, Johnson & Johnson, Kyowa Hakko Kirin Pharma, Lipid Therapeutics, LivaNova, Mallinckrodt, Medahead, MedImmune, Millenium, Mitsubishi Tanabe Pharma Corporation, MD, Nash Pharmaceuticals, Nestle, Nippon Kayaku, Novartis, Ocera, OMass, Otsuka, Parexel, PDL, Periconsulting, Pharmacosmos, Philip Morris Institute, Pfizer, Procter & Gamble, Prometheus, Protagonist, Provention, Robarts Clinical Trial, Sandoz, Schering-Plough, Second Genome, Seres Therapeutics, Setpointmedical, Sigmoid, Sublimity, Takeda, Therakos, Theravance, Tigenix, UCB, Vifor, Zealand, Zyngenia, and 4SC, as an advisory board member for Abbott Laboratories, Abbvie, Aesca, Amgen, AM Pharma, Astellas, Astra Zeneca, Avaxia, Biogen IDEC, Boehringer-Ingelheim, Bristol-Myers Squibb, Cellerix, Chemocentryx, Celgene, Centocor, Celltrion, Danone Austria, DSM, Elan, Ferring, Galapagos, Genentech, Grünenthal, Inova, Janssen, Johnson & Johnson, Kyowa Hakko Kirin Pharma, Lipid Therapeutics, MedImmune, Millenium, Mitsubishi Tanabe Pharma Corporation, MD, Nestle, Novartis, Ocera, Otsuka, PDL, Pharmacosmos, Pfizer, Procter & Gamble, Prometheus, Sandoz, Schering-Plough, Second Genome, Setpointmedical, Takeda, Therakos, Tigenix, UCB, Zealand, Zyngenia, and 4SC, and has received research funding from Abbott Laboratories, Abbvie, Aesca, Centocor, Falk Pharma, Immundiagnsotik, MD.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Klaus Gottlieb http://orcid.org/0000-0002-3747-2541
Harris Ahmad http://orcid.org/0000-0002-0890-2071
Ryan William Stidham http://orcid.org/0000-0001-9638-2186
Walter Reinisch http://orcid.org/0000-0002-2088-091X

REFERENCES

1 Abreu MT, Travis SPL, Cooney RM, *et al*. Conduct of clinical trials in Uc: impact of independent scoring of endoscopic severity on results of a randomised controlled trial: 1097. *Am J Gastroenterol* 2006;101:S429.
2 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;145:149–57.
3 Feagan BG, Vermeire S, Sandborn WJ, *et al*. Tofacitinib for maintenance therapy in patients with active ulcerative colitis in the phase 3 OCTAVE sustain trial: results by local and central endoscopic assessments. *Am J Gastroenterol* 2017;112:S329–30.
4 Ahmad H, Berzin TM, Yu HJ, *et al*. Central endoscopy reads in inflammatory bowel disease clinical trials: the role of the imaging core lab. *Gastroenterol Rep* 2014;2:201–6.
5 Johnson DA, Barkun AN, Cohen LB, *et al*. Optimizing adequacy of bowel cleansing for colonoscopy: recommendations from the US Multi-Society Task force on colorectal cancer. *Am J Gastroenterol* 2014;109:1528–45.
6 Radaelli F, Paggi S, Repici A, *et al*. Barriers against split-dose bowel preparation for colonoscopy. *Gut* 2017;66:1428–33.
7 Andrealli A, Paggi S, Amato A, *et al*. Educational strategies for colonoscopy bowel PreP overcome barriers against split-dosing: a randomized controlled trial. *United European Gastroenterol J* 2018;6:283–9.
8 Gu P, Lew D, Oh SJ, *et al*. Comparing the real-world effectiveness of competing colonoscopy preparations: results of a prospective trial. *Am J Gastroenterol* 2019;114:305–14.
9 Lawrance IC, Willert RP, Murray K. Bowel cleansing for colonoscopy: prospective randomized assessment of efficacy and of induced mucosal abnormality with three preparation agents. *Endoscopy* 2011;43:412–8.
10 Kato J, Kuriyama M, Hiraoka S, *et al*. Is sigmoidoscopy sufficient for evaluating inflammatory status of ulcerative colitis patients? *J Gastroenterol Hepatol* 2011;26:683–7.
11 Colombel J-F, Ordás I, Ullman T, *et al*. Agreement between rectosigmoidoscopy and colonoscopy analyses of disease activity and healing in patients with ulcerative colitis. *Gastroenterology* 2016;150:389–95.
12 Ringel Y, Dalton CB, Brandt LJ, *et al*. Flexible sigmoidoscopy: the patients' perception. *Gastrointest Endosc* 2002;55:315–20.
13 Hundorfean G, Pereira SP, Karstensen JG, *et al*. Modern endoscopic imaging in diagnosis and surveillance of inflammatory bowel disease patients. *Gastroenterol Res Pract* 2018;2018:5738068
14 Shaukat A, Rector TS, Church TR, *et al*. Longer withdrawal time is associated with a reduced incidence of interval cancer after screening colonoscopy. *Gastroenterology* 2015;149:952–7.
15 Rex DK, Schoenfeld PS, Cohen J, *et al*. Quality indicators for colonoscopy. *Am J Gastroenterol* 2015;110:72–90.
16 Gottlieb K, Hussain F. Voting for image scoring and assessment (VISA)--theory and application of a 2 + 1 reader algorithm to improve accuracy of imaging endpoints in clinical trials. *BMC Med Imaging* 2015;15. doi:10.1186/s12880-015-0049-0. [Epub ahead of print: Available from]. [Internet].
17 Reinisch W, Mishkin DS, Oh YS, *et al*. Impact of various central endoscopy reading models on treatment outcome in Crohn's disease using data from the randomized, controlled, exploratory cohort arm of the bergamot trial. *Gastrointest Endosc* 2020:S0016-5107(20)34355-8.
18 Harris MS, Wichary J, Zadnik M, *et al*. Competition for clinical trials in inflammatory bowel diseases. *Gastroenterology* 2019;157:S0016508519412407
19 Khanna R, Nelson SA, Feagan BG, *et al*. Endoscopic scoring indices for evaluation of disease activity in Crohn's disease. *Cochrane Database Syst Rev* 2016;8:CD010642.
20 Mohammed Vashist N, Samaan M, Mosli MH, *et al*. Endoscopic scoring indices for evaluation of disease activity in ulcerative colitis. *Cochrane Database Syst Rev* 2018;1:CD011450.
21 Limdi JK, Picco M, Farraye FA. A review of endoscopic scoring systems and their importance in a treat-to-target approach in inflammatory bowel disease (with videos). *Gastrointest Endosc* 2020;91:733–45.
22 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;13:273–84.
23 Reinisch W, Gottlieb K, Colombel J-F, *et al*. Comparison of the EMA and FDA guidelines on ulcerative colitis drug development. *Clin Gastroenterol Hepatol* 2019;17:1673–9.
24 Baron JH, Connell AM, Lennard-Jones JE. Variation between observers in describing mucosal appearances in proctocolitis. *BMJ* 1964;1:89–92.
25 Schroeder KW, Tremaine WJ, Ilstrup DM. Coated oral 5-aminosalicylic acid therapy for mildly to moderately active ulcerative colitis. A randomized study. *N Engl J Med* 1987;317:1625–9.
26 Laharie D, Filippi J, Roblin X, *et al*. Impact of mucosal healing on long-term outcomes in ulcerative colitis treated with infliximab: a multicenter experience. *Aliment Pharmacol Ther* 2013;37:998–1004.
27 Travis SPL, Schnell D, Krzeski P, *et al*. Developing an instrument to assess the endoscopic severity of ulcerative colitis: the ulcerative colitis endoscopic index of severity (UCEIS). *Gut* 2012;61:535–42.
28 Travis SPL, Schnell D, Krzeski P, *et al*. Reliability and initial validation of the ulcerative colitis endoscopic index of severity. *Gastroenterology* 2013;145:987–95.
29 Arai M, Naganuma M, Sugimoto S, *et al*. The ulcerative colitis endoscopic index of severity is useful to predict medium- to long-term prognosis in ulcerative colitis patients with clinical remission. *J Crohns Colitis* 2016;10:1303–9.
30 Ikeya K, Hanai H, Sugimoto K, *et al*. The ulcerative colitis endoscopic index of severity more accurately reflects clinical outcomes and long-term prognosis than the Mayo endoscopic score. *J Crohns Colitis* 2016;10:286–95.
31 Yarur AJ, Jairath V, Zhang J, *et al*. Tu1745 – correlation of fecal calprotectin and C-reactive protein concentrations with clinical outcomes and endoscopic disease activity in patients with ulcerative colitis receiving induction therapy with Etrasimod. *Gastroenterology* 2019;156:S-1108–S-1109.
32 Lobatón T, Bessissow T, De Hertogh G, *et al*. The modified Mayo endoscopic score (MMES): a new index for the assessment of extension and severity of endoscopic activity in ulcerative colitis patients. *J Crohns Colitis* 2015;9:846–52.
33 Rowan CR, Cullen G, Mulcahy HE, *et al*. DUBLIN [Degree of Ulcerative colitis Burden of Luminal Inflammation] Score, a Simple Method to Quantify Inflammatory Burden in Ulcerative Colitis. *J Crohns Colitis* 2019;13:1365–71.
34 Mary JY, Modigliani R. Development and validation of an endoscopic index of the severity for Crohn's disease: a prospective multicentre study. Groupe d'Etudes Thérapeutiques des Affections Inflammatoires Du tube Digestif (GETAID). *Gut* 1989;30:983–9.
35 Daperno M, D'Haens G, Van Assche G, *et al*. Development and validation of a new, simplified endoscopic activity score for Crohn's disease: the SES-CD. *Gastrointest Endosc* 2004;60:505–12.
36 Khanna R, Zou G, D'Haens G, *et al*. Reliability among central readers in the evaluation of endoscopic findings from patients with Crohn's disease. *Gut* 2016;65:1119–25.

37 Peyrin-Biroulet L, Sandborn W, Sands BE, *et al*. Selecting therapeutic targets in inflammatory bowel disease (STRIDE): determining therapeutic goals for Treat-to-Target. *Am J Gastroenterol* 2015;110:1324–38.

38 Vuitton L, Marteau P, Sandborn WJ, *et al*. IOIBD technical review on endoscopic indices for Crohn's disease clinical trials. *Gut* 2016;65:1447–55.

39 Ferrante M, Colombel J-F, Sandborn WJ, *et al*. Validation of endoscopic activity scores in patients with Crohn's disease based on a post hoc analysis of data from sonic. *Gastroenterology* 2013;145:978–86.

40 Khanna R, Zou G, Stitt L, *et al*. Responsiveness of endoscopic indices of disease activity for Crohn's disease. *Am J Gastroenterol* 2017;112:1584–92.

41 Rutgeerts P, Geboes K, Vantrappen G, *et al*. Predictability of the postoperative course of Crohn's disease. *Gastroenterology* 1990;99:956–63.

42 de Jong MJ, Huibregtse R, Masclee AAM, *et al*. Patient-Reported outcome measures for use in clinical trials and clinical practice in inflammatory bowel diseases: a systematic review. *Clin Gastroenterol Hepatol* 2018;16:648–63.

43 Gracie DJ, Williams CJM, Sood R, *et al*. Poor correlation between clinical disease activity and mucosal inflammation, and the role of psychological comorbidity, in inflammatory bowel disease. *Am J Gastroenterol* 2016;111:541–51.

44 Kane SV, Chandrasekhara V, Sedlack RE, *et al*. Credentialing for endoscopic practice: the Mayo clinic model. *Clin Gastroenterol Hepatol* 2018;16:1370–3.

45 Rex DK, Petrini JL, Baron TH, *et al*. Quality indicators for colonoscopy. *Gastrointest Endosc* 2006;63:S16–28.

46 Ezaz G, Leffler DA, Beach S, *et al*. Association between endoscopist personality and rate of adenoma detection. *Clin Gastroenterol Hepatol* 2019;17:1571–9.

47 Fernandes SR, Pinto JSLD, Marques da Costa P, *et al*. Disagreement among Gastroenterologists using the Mayo and Rutgeerts endoscopic scores. *Inflamm Bowel Dis* 2018;24:254–60.

48 Daperno M, Comberlato M, Bossa F, *et al*. Training programs on endoscopic scoring systems for inflammatory bowel disease lead to a significant increase in interobserver agreement among community Gastroenterologists. *J Crohns Colitis* 2017;11:556–61.

49 Strickland NH. Quality assurance in radiology: peer review and peer feedback. *Clin Radiol* 2015;70:1158–64.

50 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.

51 Gottlieb K, Travis S, Feagan B, *et al*. Central reading of endoscopy endpoints in inflammatory bowel disease trials. *Inflamm Bowel Dis* 2015;21:1.

52 Central limit theorem - Encyclopedia of Mathematics [Internet], 2019. Available: https://www.encyclopediaofmath.org/index.php/Central_limit_theorem

53 Liddell TM, Kruschke JK. Analyzing ordinal data with metric models: what could possibly go wrong? *J Exp Soc Psychol* 2018;79:328–48.

54 Ahmad HA, Gottlieb K, Hussain F. The 2 + 1 paradigm: an efficient algorithm for central reading of Mayo endoscopic subscores in global multicenter phase 3 ulcerative colitis clinical trials. *Gastroenterol Rep* 2016;4:35–8.

55 Reeve R, Gottlieb K. Sequentially determined measures of interobserver agreement (kappa) in clinical trials may vary independent of changes in observer performance. *Ther Innov Regul Sci* 2020;54:681–6.

56 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.