Establishment of a validated central reading system for ileocolonoscopy in an academic setting

Tim Raine, Holly Pavey, Wendi Qian, Gordon W Moran, Sreedhar Subramanian, Lizzie Swaby, Simon PL Travis, Shahida Din, Peter M Irving, James O Lindsay, Miles Parkes, Nicholas A Kennedy

BACKGROUND
Therapeutic interventions in IBD should alleviate patient symptoms and modify long-term disease outcomes. This second target is hard to demonstrate within the context of typical observational or interventional studies due to problems of sample size and duration of follow-up. Instead, there has been considerable focus on the assessment of mucosal inflammation using validated scoring systems in both Crohn’s disease (CD) and UC. Achievement of endoscopic healing correlates with avoidance of long-term disease complications in both CD and UC, and is an established therapeutic target in addition to the achievement of clinical remission. Therapeutic interventions in IBD should aim to achieve endoscopic healing in a proportion of patients.

For these reasons, data on endoscopic outcomes are now required by regulatory agencies for any registration trial. In addition, endoscopic outcome data may be used in earlier phase trials to guide decisions about further asset development, adding biological context to clinical data. Interventions which lead to improvements in symptom scores may not demonstrate improvements in endoscopic outcomes and this additional information may inform or limit placement in treatment paradigms. Furthermore, the use of endoscopic activity as an inclusion criterion for clinical trials can reduce the high placebo response rates previously reported in trials reliant purely on clinical inclusion criteria, which may mask efficacy signals.

Key barriers to use of endoscopy in clinical trials include patient acceptability. Additionally, there is a need for blinded central reading of all trial endoscopies to ensure standardisation and avoid inadvertent introduction of investigator bias: local endoscopy reads are typically less reliable than blinded central reading by a recognised expert reader. Reliance on local readers may even lead to critical errors in trial interpretation. Furthermore, use of central readers alone does not completely mitigate against errors of endoscopy reading that may lead to false conclusions about novel interventions, underscoring the importance of ensuring validity and quality in any reader pool.

Currently, access to a validated pool of expert readers may be beyond the budget of typical investigator-initiated studies or early phase industry studies, which may instead rely on local reads or unvalidated central readers. Previous reports have demonstrated the validity and reliability of small expert reader pools. Nevertheless, the methodology and feasibility of undertaking the process of establishing a validated pool of central readers has not previously been reported. We developed a protocol and platform to support the validation of a pool of central readers, using the simple endoscopic score for CD (SES-CD) as an example. This is a relatively straightforward and reliable score that shows good interobserver and intra-observer variability after training. We report here the methodology and the results of this exercise as well as a process for maintaining and assuring the quality of the pool of readers for future studies.

METHODS AND RESULTS

Platform
We developed a platform to enable uploading, processing and reading of endoscopic videos using free-to-use software. The core web-based system was implemented using C# and ASPNET core (.NET V5.0; Microsoft, Redmond, USA), with data stored in MySQL 8 (Oracle, Redwood Shores, USA). Minimum video resolution was 720×480 pixels. Video playback was provided for central readers using Video.js (Brightcove Inc, Boston, USA).

Initial establishment of a reader pool
A library of anonymised ileocolonoscopy videos from patients with CD was created on the platform. Fully anonymised videos were used with appropriate informed consent obtained at the time of index examination using internal institutional approvals. Videos were acquired by endoscopists under routine clinical conditions; endoscopists were made aware that the videos were being recorded for scoring purposes and reminded to ensure adequate mucosal visualisation, but not provided with any further instructions or specific training. Endoscopists were requested not to upload videos with bowel preparation judged to be inadequate for mucosal assessment. The platform provides the option for readers to indicate that bowel preparation was inadequate for scoring purposes, but this option was not used by any of the readers in the present exercise.

Video scoring and analysis proceeded according to a prespecified analysis plan. Seven prospective scorers were identified from a group of experienced clinical trialists, all with prior experience of use of the SES-CD as local readers. Scorers were invited to attend an initial videoconference where the SES-CD was reviewed and the use of the platform demonstrated. In particular, areas of common difficulty were discussed and a series of ‘Top Tips’ agreed and provided as an aid for reference during subsequent scoring (see online supplementary material 1). Although one reader acted as chair and facilitated the
discussion (IL), no one individual acted as a ‘trainer’ but rather the group discussed use of the score, as well as reviewing a series of endoscopic videos selected to illustrate a range of commonly encountered challenges.

The scorers were then asked individually to score sets of videos taken in random batches of 10 from the library (without replacement), using a standardised template to report the SES-CD score and subcomponents (online supplemental material 1). Inter-rater reliability scores were determined by a blinded independent assessor using intraclass correlation coefficients (ICCs) and 95% CI, with a two-way mixed-effects model (treating readers as fixed and videos as random variables). We wanted the eventual pool of readers to allow a high degree of confidence in the absolute score assigned by any single reader chosen at random from within this fixed pool, hence used an ICC calculation based on ‘single reader, absolute agreement’. The final formula used for the ICC calculation was:

$$\text{ICC} = \frac{\text{MS}_R - \text{MS}_E}{\text{MS}_R + (k-1)\text{MS}_E + \frac{k}{n} (\text{MS}_C - \text{MS}_E)}$$

where MS R denotes mean square for rows; MS E denotes mean square for error; MS C = mean square for columns; n = number of videos; k = number of readers. Analysis was performed using R V4.0.21 with the ICC calculated using the irr package. As a general rule, ICC values of >0.6 can be regarded as good and >0.75 as excellent.\(^{22}\)

### Video scoring and analysis plan

According to the prespecified plan, should the initial estimate of the ICC for the first set of 10 videos be <0.7, then these scores would be removed from further rounds of ICC calculation and the first 10 videos regarded as a training exercise. In our study, the ICC after the first 10 videos was 0.77 (95% CI: 0.56 to 0.92), so these scores were retained in subsequent rounds of analysis.

In addition, the ICC was recalculated dropping each scorer in turn to assess any scorer whose inclusion adversely affected the performance of the group—we did not identify any such outliers. ICC were calculated for each subscore of the SES-CD, with the lowest ICC identified for the stenosis subscore of the SES-CD. These findings were all fed back to the scorers at a further videoconference, at which points of discrepancy were discussed and videos of segments showing high inter scorer variability reviewed.

Next, further batches of 10 randomly selected videos were scored, with the ICC calculations run after each set. We prespecified that the scoring exercise would be terminated when any of the following criteria were met:

1. ICC >0.7 with 95% CI for ICC >0.6
2. Exhaustion of available videos in library for scoring
3. Agreement among scorers that exercise should be terminated.

Table 1 shows the estimated number of videos required for k readers scoring for expected ICC’s in the range 0.70 to 0.85. Sample size based on a two-way mixed-effects model. Note that LCI represents the lower limit of 2-way 95% CI - that is, value below which the true ICC of the sample of readers is <2.5% likely to fall. ICC, intraclass correlation coefficient; LCI, lower limit of two-way 95% CI.

In fact, after a total of 20 videos had been scored, the ICC was 0.82 (95% CI: 0.71 to 0.91) which meant that we terminated the exercise without needing further scoring rounds. No single reader significantly adversely affected the performance of the pool (ICC estimates remained in range 0.80–0.83 when each reader was removed). As determined after completion of the scoring exercise, the median SES-CD score was 7 (IQR: 4–11). Table 2 shows further characteristics of the videos used. Notably, despite the modest number of videos ultimately read, a wide range of disease severities was represented. The maximum SES-CD score encountered was 22, while one video had an SES-CD of 0. One video had an impassable stricture, while 10 videos included patients with prior resectional surgery. In three patients, the extent of endoscopic examination was less than complete (with a combined total of five segments of unexplored bowel).
Ongoing quality assurance

The performance of readers may change over time. To assess this, we invited readers to rescore a selection of videos after a period of at least 6 months since the original reading and without any further training. Videos were selected and allocated to readers at random, with subsequent adjustment to ensure that each video was rescored by at least one reader. Readers were asked to rescore three videos, based on a prior power calculation assuming 80% power to detect an intrarater ICC score of 0.8 (online supplemental material 2). Using a two-way, mixed-effects, single rater, absolute agreement approach, we determined the intrarater ICC as 0.937 (95% CI: 0.877 to 0.968).

To monitor and assure quality, our platform automatically assigns a minimum of 20% of all videos scored on the platform are automatically assigned to a pair of randomly selected readers, both blinded to this. After each reader has completed at least 20 such paired readings, the ICC is recalculated using a two-way random-effects model applied to all paired readings available since the last performance check. In the event that the lower limit of the 95% CI for the ICC decreases to ≤0.6, the ICC is then recalculated with the sequential exclusion of each reader in the pool to establish whether the exclusion of any one individual lead to an improvement in the lower limit of the 95% CI to >0.6. If an underperforming reader is identified as contributing to the change in reliability, this reader would be temporarily removed from the pool and offered further training. Parallel reading continuing prospectively for a further 20 videos—at which point ICC would be recalculated using these scores and the same performance criteria applied.

New readers

For new readers wishing to join the pool, it is prespecified that the reader would be offered training, then invited to read the original 10 videos in the training exercise. Their scores for these videos would then be used to calculate an ICC using the new reader’s scores in the place of each reader in turn from the original group of seven readers. The process specifies that the ICC should not decrease with more than three of the combinations applied, meaning that each new reader would need to show consistency in line with the median of the original seven readers. To test this process, we invited an eighth reader to score the original 10 videos and assessed their read quality in this manner and found that when the new reader was substituted for three of the original seven, the ICC decreased (lowest value was 0.73), while in the other four substitutions, the ICC increased (highest value was 0.79). The new reader was therefore admitted to the pool.

DISCUSSION

Central reading of endoscopy videos has played an ever-increasing role in clinical trial conduct, but to date there have been no reports on the protocol for, or establishment of, a validated pool of central readers suitable for use in the context of investigator-initiated studies. Khanna et al reported the detailed analysis of expert reader performance using a range of CD endoscopy scores and showed good performance of the SES-CD when using four expert central readers. The overall intrarater ICC reported for SES-CD scores in this study was 0.83 (0.75-0.88), against which our reported ICC of 0.82 (95% CI: 0.71 to 0.91) compares favourably, although it must be noted that Khanna et al used a slightly different approach to ICC derivation due to the different methodology in their study. While a high ICC reflects strong agreement between readers, a note of caution common to all such studies is the absence of a defined ‘gold standard’ against which to measure overall performance of the reader pool and exclude the possibility of a systematic error common to all readers.

Previous studies have examined how non-expert readers can be trained, leading to improvements in interobserver agreement (though with wider CIs around the ICC than those we report here). However, until the present study the method and means to establish a central read pool and the feasibility of doing so, using readers without the same degree of prior experience in central reading, had not been reported. In our study, after just 20 videos and 2 teleconferences, our 7 readers were able to demonstrate a high degree of reliability and an eighth reader could be admitted to the pool after similar training. Indeed, the termination of the scoring exercise after just 20 videos likely reflects the importance of adequate training prior to the exercise for a group of readers with prior experience of using the score, and the prior opportunity for discussion and consensus around areas of common difficulty.

We calculated an SES-CD metric that reflects the confidence that can be placed in any single read made by a reader randomly selected from the pool. We also demonstrate how the pool can be added to and monitored for quality assurance over time. There are arguments in favour of using more than a single reader—either using statistical methods to combine reader scores and detect outliers or via methods of arbitration. While this may lead to further increases in confidence in scores, and is supported within our platform for future users who may require this level of confidence, this is not an absolute requirement of our system. Importantly, a minimum percentage of reads must be double-read to support ongoing quality assurance of the reader pool—a process that is implemented in our platform in a blinded manner.

While we have developed and validated our system to work with SES-CD scores, the methodological principles we set out here might be equally applicable to other scoring systems where the outcome can be handled as a continuous variable. These include not only endoscopic scores in UC, but also other systems where human raters are asked to assess a visual image against a validated scoring system, such as occurs, for example, in the scoring of MRI. Nevertheless, other scoring systems might differ in complexity, require more extensive training and necessitate much larger image sets and more iterations of the scoring and training process during any validation exercise.

This exercise shows that it is feasible to establish a validated pool of central readers to support clinical trials using individuals with previous experience of use of the scoring system. This pool of readers and the supporting platform is available to support investigator-initiated clinical
trials, on a not-for-profit basis. Our methodology is reproducible by other investigators seeking to establish their own validated reader pools.

Acknowledgements SD acknowledges the support of NHS Research Scotland through NHS Lothian. JL acknowledges the support of the NIHR EME programme.

Contributors Study conception and design was done by TR, HP, WQ, JL, MP and NAK. Study was conducted by TR, HP, ES, JL, MP and NAK. Data generation was done by GWM, SS, SPLIT, SD, PMI, JL, MP and NAK. Analysis and reporting was performed by TR, HP, JL, MP and NAK. All authors were involved in manuscript writing and gave approval of the final version.

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 None declared.

Patient consent for publication Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Supplemental material This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and content. Where the content includes any translated material, BMJ does not endorse by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not endorse.

© Author(s) (or their employer(s)) 2022. No commercial re-use. See rights and permissions. Published by BMJ.

► Additional supplemental material is published online. To view, please visit the journal online (http://dx.doi.org/10.1136/gutjnl-2021-325575).

To cite Raine T, Pavey H, Qian W, et al. Gut 2022;71:661–664. Received 5 July 2021 Accepted 23 December 2021 Published Online First 12 January 2022 Gut 2022;71:661–664, doi:10.1136/gutjnl-2021-325575

ORCID IDs Tim Raine http://orcid.org/0000-0002-5855-9873

Gordon W Moran http://orcid.org/0000-0002-8906-6908

Sreedhar Subramanian http://orcid.org/0000-0002-6483-1730

Simon PL Travis http://orcid.org/0000-0002-2690-4361

Shahida Din http://orcid.org/0000-0003-2855-3400

Peter M Irving http://orcid.org/0000-0003-0972-8148

James O Lindsay http://orcid.org/0000-0003-3353-9590

Miles Parke http://orcid.org/0000-0002-6467-0631

Nicholas A Kennedy http://orcid.org/0000-0003-4368-1961

REFERENCES

1 Mary JT, 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.

2 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.

3 Schroeder KW, Tremaine WJ, stump 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.

4 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.

5 Fretste KE, Jahnson J, Moum BA, et al. Mucosal healing in inflammatory bowel disease: results from a Norwegian population-based cohort. Gastroenterology 2007;133:412–22.

6 Shah SC, Colombel J-F, Sands BE, et al. Systematic review with meta-analysis: mucosal healing is associated with improved long-term outcomes in Crohn’s disease. Aliment Pharmacol Ther 2016;43:317–33.

7 Turner D, Ricciuto A, Lewis A, et al. STRIDE-II: an update on the selecting therapeutic targets in inflammatory bowel disease (STRIDE) initiative of the International organization for the study of IBD (IOIBD): determining therapeutic goals for Treat-to-Target strategies in IBD. Gastroenterology 2016;150:750–83.

8 Reinsch 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.

9 Feagan BG, Sands BE, Rossiter G, et al. Effects of Mongersen (GED-0301) on endoscopic and clinical outcomes in patients with active Crohn’s disease. Gastroenterology 2018;154:61–6.

10 Nafzati T, Bar-Lev Schneider L, Sickerovsky Benjamin Y, et al. Cannabis is associated with clinical but not endoscopic remission in ulcerative colitis: a randomized controlled trial. PloS One 2021;16:e0246871.

11 Graham DY, Naser SA, Offman E, et al. 643 RH-104, a Fixed-Dose, Oral Antibiotic Combination Against Mycobacterium Avium Paratuberculosis (MAP) Infection, Is Effective in Moderately to Severely Active Crohn’s Disease. Am J Gastroenterol 2019;114:5376–7.

12 Reinsch W, de Villiers W, Bene L, et al. Fontolizumab in moderate to severe Crohn’s disease: a phase 2, randomized, double-blind, placebo-controlled, multiple-dose study. Inflamm Bowel Dis 2010;16:233–42.

13 Harris MS, Wichary J, Zadnik M, et al. Competition for clinical trials in inflammatory bowel diseases. Gastroenterology 2019;157:1457–61.

14 Hébuterne X, Lémann M, Bouhnik Y, et al. Endoscopic improvement of mucosal lesions in patients with moderate to severe ileocolonic Crohn’s disease following treatment with certolizumab pegol. Gut 2013;62:201–8.

15 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.

16 Sandborn WJ, Bressler B, Lee S, PTG-100, an oral gut-restricted peptide celebrity antagonist, induces clinical and histologic remission in patients with moderate to severely active ulcerative colitis. United Eur Gastroenterol J 2018.

17 Hawkey CI, Allez M, Clark MM, et al. Autologous hematopoietic stem cell transplantation for refractory Crohn disease: a randomized clinical trial. JAMA 2015;314:2524–34.

18 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.

19 Shoult PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86:420–8.

20 Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J Clin Exp Med 2015;1:155–63.

21 R Core Team. R: a language and environment for statistical computing. Vienna, Austria: R A Lang. Environ. Stat. Comput. R Found. Stat. Comput, 2020.

22 Cicchetti DV. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychol Assess 1994;6:284–90.

23 Daperno M, Comberato 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.

24 Panés J, Feagan BG, Hussain F, et al. Central endoscopy reading in inflammatory bowel diseases. J Crohns Colitis Colitis 2016;10 Suppl 2:S52–7.

25 Gottlieb K, Daperno M, Usukin K, et al. Endoscopy and central reading in inflammatory bowel disease clinical trials: achievements, challenges and future developments. Gut 2021;70:418–26.