Journal of Crohn's and Colitis, 2017, 556–561
doi:10.1093/ecco-jcc/jjw181
Advance Access publication October 4, 2016
Original Article

Original Article

Training Programs on Endoscopic Scoring Systems for Inflammatory Bowel Disease Lead to a Significant Increase in Interobserver Agreement Among Community Gastroenterologists

Marco Daperno[a], Michele Comberlato[b], Fabrizio Bossa[c], Alessandro Armuzzi[d], Livia Biancone[e], Andrea G Bonanomi[f], Andrea Cassinotti[g], Rocco Cosintino[h], Giovanni Lombardi[i], Roberto Mangiarotti[h], Alfredo Papa[d], Roberta Pica[j], Luca Grassano[k], Guido Pagana[k,l], Renata D'Incà[m], Ambrogio Orlando[n], Fernando Rizzello[o], on behalf of the IGIBDEndo Group

[a]Gastroenterology Unit, AO Ordine Mauriziano, Torino, TO, Italy [b]Gastroenterology Unit, Ospedale di Bolzano, Bolzano, Italy [c]Gastroenterology Unit, IRCCS 'Casa Sollievo della Sofferenza', San Giovanni Rotondo, Italy [d]Gastroenterology Unit, Complesso integrato Columbus, Roma, Italy [e]Gastroenterology Unit, Tor Vergata University, Roma, Italy [f]Gastroenterology Unit, AOU Careggi, Firenze, Italy [g]Gastroenterology and IBD Unit, "Luigi Sacco" University Hospital, Milan, Italy [h]Gastroenterology Unit, S. Camillo-Forlanini Hospital, Roma, Italy [i]Gastroenterology Unit, Cardarelli Hospital, Napoli, Italy [j]Gastroenterology Unit, ASL Roma B, Ospedale Pertini, Roma, Italy [k]Politecnico di Torino, Torino, Italy [l]Istituto Mario Boella, Torino, Italy [m]Department of Surgery, Oncology and Gastroenterology, Azienda Ospedaliera di Padova, Padova, Italy [n]Internal Medicine Unit, AO Ospedali Riuniti Villa Sofia – Cervello, Palermo, Italy [o]Internal Medicine Unit, Policlinic S. Orsola Malpighi and Bologna University, Bologna, Italy

Corresponding author: Marco Daperno, MD, Gastroenterology Unit, AO Ordine Mauriziano, Largo Turati 62, I-10128 Torino, Italy. Tel: +39-011-508-2534; Fax: +39-011-5082-536; Email: mdaperno@teletu.it

Abstract

Background and Aims: Endoscopic outcomes are increasingly used in clinical trials and in routine practice for inflammatory bowel disease [IBD] in order to reach more objective patient evaluations than possible using only clinical features. However, reproducibility of endoscopic scoring systems used to categorize endoscopic activity has been reported to be suboptimal.
The aim of this study was to analyse the inter-rated agreement of non-dedicated gastroenterologists on IBD endoscopic scoring systems, and to explore the effects of a dedicated training programme on agreement.
Methods: A total of 237 physicians attended training courses on IBD endoscopic scoring systems, and they independently scored a set of IBD endoscopic videos for ulcerative colitis [with Mayo endoscopic subscore], post-operative Crohn's disease [with Rutgeerts score] and luminal Crohn's disease (with the Simple Endoscopic Score for Crohn's Disease [SESCD] and Crohn's Endoscopic Index of Severity [CDEIS]). A second round of scoring was collected after discussion about

Abbreviations: SESCD: Simple Endoscopic Score for Crohn's Disease; CDEIS: Crohn's Endoscopic Index of Severity; Kappa: kappa statistics; IGIBD: Italian group for inflammatory bowel disease.

determinants of discrepancy. Interobserver agreement was measured by means of the Fleiss' kappa [kappa] or intraclass correlation coefficient [ICC] as appropriate.
Results: The inter-rater agreement increased from kappa 0.51 (95% confidence interval [95% CI] 0.48–0.55) to 0.76 [95% CI 0.72–0.79] for the Mayo endoscopic subscore, and from 0.45 [95% CI 0.40–0.50] to 0.79 [0.74–0.83] for the Rutgeerts score before and after the training programme, respectively, and both differences were significant [$P < 0.0001$]. The ICC was 0.77 [95% CI 0.56–0.96] for SESCD and 0.76 [0.54– 0.96] for CDEIS, respectively, with only one measurement.
Discussion: The basal inter-rater agreement of inexperienced gastroenterologists focused on IBD management is moderate; however, a dedicated training programme can significantly impact on inter-rater agreement, increasing it to levels expected among expert central reviewers.

Key Words: Endoscopic scoring; inter-rater agreement; teaching; Crohn's disease; Mayo endoscopic subscore; Rutgeerts score; ulcerative colitis

1. Background

Measuring endoscopic activity in inflammatory bowel disease [IBD] has attracted increasing interest, as it has been shown that at least in clinical trials[1,2] there is a substantial and relevant difference between local and central reviewers' scores—which might lead to slightly [but significantly] different results.[1] Moreover, since purely clinical endpoints are suboptimal for assessing deep and durable remission,[3–6] and since in some cases clinical scores have been shown not to correlate significantly with biochemical surrogate endpoints or with endoscopic activity,[7] measurement of endoscopic activity has been proposed as a pertinent and more objective endpoint for future clinical trials and for clinical practice.[8]

However, many studies over the past 5 years have pointed out that intra- and inter-observer agreement might be suboptimal for a number of endoscopic scoring systems,[1,9–14] especially where ulcerative colitis is concerned.[1,9–11] The agreement performance between a few highly experienced endoscopists with considerable skills in central review has been shown to be extremely good;[1,12,14] however, there is evidence that unexperienced clinicians might have significantly less agreement performance.[9,13]

One solution to this issue, which may be considered only for clinical trials and which generates increasing costs for guaranteeing high quality of referrals, is to adopt central review systems.[1,8,15] Alternatively, if a shared learning process could result in similar performance agreement results, it is proposed that local endoscopists could reach adequate levels of proficiency in scoring IBD endoscopy, and that they might achieve levels of inter-observer agreement comparable with those displayed by expert central reviewers, at much lower cost.

The aim of this study was to verify whether a learning project dedicated to IBD endoscopic scoring had a significant impact on inter-observer agreement.

2. Materials and methods

2.1. Learning and teaching module

A learning project was carried out focused on IBD endoscopic scores, which involved 237 Italian physicians with interest in IBD management, in 14 venues [seven meetings in 2 consecutive years]. The meetings involved 25–30 participants each time, and every participant attended at least a single meeting [in 185/237 cases, 78%]; in some cases participants attended two meetings in two subsequent years. The attendees were all gastroenterologists or internists with a minimum post-certification experience of 3 years and a maximum experience of 30 years, and all were actively involved both in an IBD clinic and in endoscopy. During the meetings, attendees received information on how to score four major endoscopic scoring systems: the Mayo endoscopic subscore for ulcerative colitis,[16] the Rutgeerts score for postoperative recurrence,[17] the Crohn's disease endoscopic index of severity [CDEIS],[18] and the simple endoscopic score for Crohn's disease [SES-CD][19] for luminal Crohn's disease.

The teaching module was based on a slide-set presentation dedicated to the characteristics and interpretation of each individual scoring system [a total of four presentations], and to common pitfalls and problems in scoring. This was integrated with a selection of short endoscopic clips relevant to the descriptions of elemental lesions and the most common endoscopic patterns. The presentations were repeated in all meetings, based on the same template, and they lasted 45–60 min. The faculty members presenting the teaching modules all had a minimum of 10 years of experience in IBD management and IBD endoscopy, and were used to IBD endoscopic scores for trial reasons. Moreover, most of them had been involved in a previous study on scoring reproducibility.[13] Adequate discussion time was allocated during and at the end of each presentation, in order to engage all participants in the teaching process.

After the presentations, every attendee independently reviewed and scored with the relevant scoring system a number of endoscopic anonymized videos [between five and six in each meeting] for each scoring system. The endoscopic video library was recorded at standard definition in white light, and anonymized before collection, by members of the faculty; it was used both for development of the teaching module and [with a separate subset of videos] for testing readers' agreement. Videos shown to the audience lasted between 2 min [in some cases of ulcerative colitis or post-surgical Crohn's recurrence] and 12 min [for the longest luminal Crohn's disease video]. Each attendee reviewed the videos on an iPad [Apple Inc., USA], in order to guarantee independency of evaluation and scoring, and the chance for each individual attendee to review videos at their preference, and at an adequate resolution. At the end of each visualization, voting of every participant was forced, before stepping forward to subsequent videos. Votes were collected by means of a dedicated software, slightly modified for networking and recording needs, to the IGIBD Scores App [http://www.igibdscores.it/en/, IGIBD, Italy] before going on to subsequent videos.

After the scores were recorded on the local server, a general discussion between participants on reasons for agreement and disagreement took place. The video and discussion sessions lasted 1–2 h depending on the number and complexity of the videos and the length of the discussions, and they were chaired by faculty experts. At the end of the meeting, attendees were re-administered the same

video sets in a random fashion, and again they were required to independently score the relevant activity score [for ulcerative colitis and postoperative Crohn's disease] for each video, with the same setting.

2.2. Statistical analysis

Inter-observer agreement was measured by means of Fleiss' kappa statistics for the Mayo endoscopic subscore and for the Rutgeerts score. Since the observations of different observers were available on two different subsets of videos, a meta-analytical approach was used to summarize the data between the 2 years, studying the independent measurements and averaging the separate results.

If videos were not assessed twice [before and after training and discussion], those votes were discarded. Since some attendees participated in the two successive years of the learning project, their votes in the second year were discarded, based on the fact that they were not naïve to the learning programme. For luminal Crohn's disease [since re-evaluation of those videos required much longer times], no repeated evaluation was available, and therefore no pre-/post-training evaluation was available. Scores were referred to the individual physicians through the tablet and were recorded on a server, both for teaching and study purposes. Details of the maintained/discarded votes are reported in Figure 1.

Variations of the kappa values within the same year were tested according to a fixed raters layout analysed in a recent paper,[20] accounting for the dependence between repeated evaluations and non-homogeneous subjects. A generalization to more than two categories was needed in order to obtain results about non-dichotomous ratings.

For luminal Crohn's disease scores [CDEIS and SES-CD], an intraclass correlation coefficient with dispersion measures was used, but no pre-/post-training variation in scores was measurable due to a single time-point evaluation. In order to pool the results from the meetings in the two calendar years, a rather different approach was used considering a pooling by weighted average; however, this result is outside the focus of this paper, namely the educational effect, and the large number of observers.

3. Results

3.1. Ulcerative colitis agreement results

In the first year of the learning project, a set of five ulcerative colitis videos was observed at least once by 177 attendees; 121 were valid for complete analysis, with complete pre- and post-training scores. In the second year of the learning project, a set of six different ulcerative colitis videos was observed at least once by 155 attendees; 64 were valid for complete analysis, leading to a pool of 185 readers with complete pre- and post-training evaluations for the Mayo endoscopic subscore [Figure 1A].

Kappa values for inter-observer agreement before and after the training programme were 0.51 (95% confidence interval [95% CI] 0.48–0.55) and 0.76 [95% CI 0.72–0.79], respectively, leading to a delta-kappa of 0.24 [95% CI 0.20–0.29], with $P < 0.0001$. A graphical

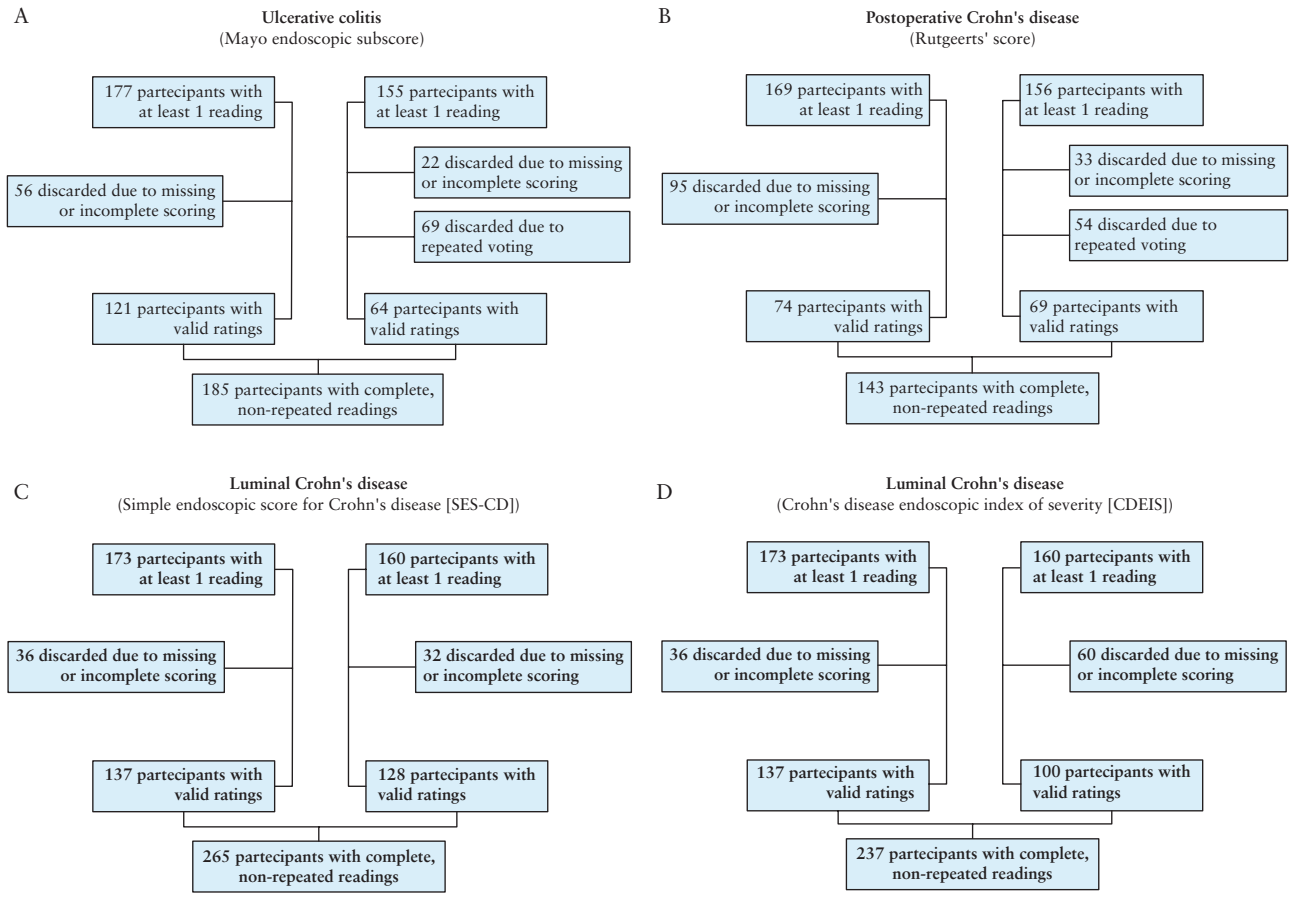

Figure 1. A–D. Graphical representation of the flow of readings in the second year of the projects, with reason to discard scores [incomplete reads or reads belonging to attendees not naïve to the learning programme]. Data for ulcerative colitis [**1A**], post-operative Crohn's disease [**1B**] and ileocolonic 'luminal' Crohn's disease [**1C–D**] are presented. CDEIS: Crohn's disease index of severity; SES-CD: simple endoscopic score for Crohn's disease.

representation of the kappa values before and after the training programme is presented in Figure 2. Individual pre- and post-training kappa values for inter-observer agreement are reported in Table 1.

3.2. Post-operative Crohn's agreement results

Regarding the post-operative videos, during the first year of the project 169 attendees reviewed a set of five videos at least once; after deleting partial or missing observations, 74 were valid for complete analysis. In the second year of the learning project, a set of 6 different pertinent videos was reviewed at least once by 156 attendees; after deleting missing cases or repeated participations, 69 were valid for complete analysis, leading to a pool of 143 readers with complete pre- and post-training evaluations for Rutgeerts score [Figure 1B].

Kappa values for inter-observer agreement before and after the training programme were 0.45 [95% CI 0.40–0.50] and 0.79 [95% CI 0.74–0.83], respectively, leading to a delta-kappa of 0.34 [95% CI 0.28–0.39], with $P < 0.0001$. Graphical representation of the kappa values before and after the training programme is presented in Figure 2. Individual pre- and post-training kappa values for inter-observer agreement are reported in Table 1.

3.3. Luminal Crohn's results

For luminal Crohn's disease, the numbers of videos scored were 5 and 6, respectively, in the 2 years of the learning project; attendees scoring the complete set of videos in the first and second year of the project were 137 and 128 for SES-CD, respectively, and 137 and 100 for CDEIS, respectively. Therefore, 265 complete readers-reads pairs for SES-CD and 237 for CDEIS, respectively, were available [Figure 1B–C]. A detailed report of votes retained and discarded, with the reasons for discarding individual observers' data, is presented in Figure 1A–D.

As mentioned, for SES-CD and CDEIS, only one round of observation was performed; therefore, only basal agreement measures are available. The values obtained in the first and second years of the project, together with the summary results, are reported in Table 1.

4. Discussion

Endoscopic scoring systems are increasingly employed for assessment of disease activity in IBD, and for tailoring therapy accordingly, even

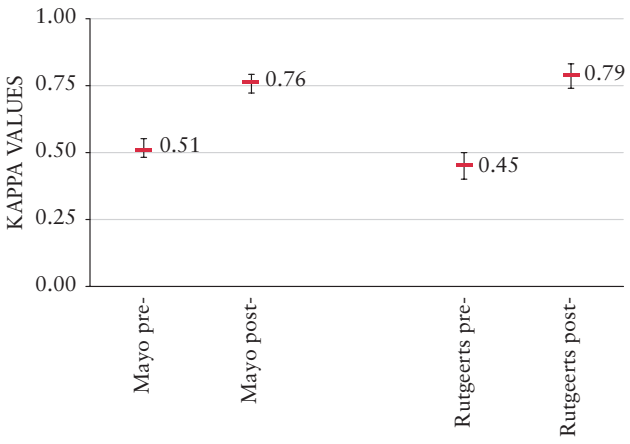

Figure 2. Graphical report of interobserver agreement results for Mayo endoscopic subscore [Mayo] and Rutgeerts score [Rutgeerts]. Results are expressed as kappa values before [pre-] and after [post-] training programme, along with dispersion measures [95% confidence intervals]. Differences for Mayo and Rutgeerts score are statistically significant [both $P < 0.0001$].

when symptoms are minimal or absent.[2–8] However, there is growing interest in exploring the reproducibility of the most commonly used endoscopic scores for inflammatory bowel disease.[1,2,9–14,18,19]

Although endoscopic activity is an objective effect of inflammation, endoscopic scores are algorithms for interpreting and describing images, which inherit a degree of variability and subjectivity due to the fact that they are interpretation of images. Discussion concerning this in the IBD field has been to date relatively limited and recent—the main method of overcoming subjectivity in the interpretation of endoscopic activity, and of guaranteeing accuracy of trial results, has been the adoption of central reviewing paradigms, with dedicated experts[1,2,6,8,12,14,15] responsible for consistent scoring throughout a trial.

For a post-operative CD patient with a Rutgeerts score of i2 and above, escalation of therapy is usually considered.[21–24] However, as reported by a recent paper using a large set of observations, reproducibility of Rutgeerts score might be poorer than previously reported,[25] especially when differentiation around the i2 grade is concerned. The volume of observations in that paper was very large [13 experts analysing a subset of 39 videos = 156 readings] and the main result was a kappa of 0.47 for distinguishing i0–1 from i2–4, and a kappa of 0.64 for distinguishing i0–2 from i3–4, while the five-grade weighted kappa was 0.43. Such results cannot be easily compared with this study results, but the volume of observations [even if on a much smaller number of videos] is by far larger [74 observers × 5 videos + 69 observers × 6 videos, leading to 784 pairs of observations]. Moreover, the overall results of the former study compare well with the basal results reported in the present study [kappa = 0.45, 95% CI 0.40–0.50], which were substantially ameliorated by the educational programme, with a significant increase in inter-observer agreement to kappa values of 0.79 [95% CI 0.74–0.83].

In this study we aimed to evaluate, as a proof of concept, the use of a dedicated training program for IBD endoscopic scores, including a peer discussion of determinants of disagreement. Unlike central reader systems, agreement among many IBD physicians, might be more reflective of 'crowd wisdom' and more suitable for real-life clinical management, and could become the keystone for applying clinical trial results to actual life.

The main limitations of our study were the limited number of observations considered [five to six videos only] and the short-term observation of the effects of the learning programme. The first issue might create an under- or overestimation of the effects. The methodological paper[20] at the base of our analysis indicates that better results in estimating standard errors are obtained when the cardinalities of observers and subjects are both large. Relative bias in this estimation was assessed through simulation, and it was proven not to have a significant impact on the scope of our study. Short-term re-evaluation of the agreement might lead to an overestimation of the learning effects—this issue requires long-term off-site re-evaluation of attendees' performance in terms of inter-observer agreement years after the initial training.

On the other hand, the number of observers that completed the training program is the largest considered to date, and it represents a nice simulation of real-world practice. To date, one only study has analysed the performance of a dedicated training programme on the proficiency of gastroenterologists inexperienced in SES-CD, and it found consistent results of amelioration of inter-observer agreement for five gastroenterologists. The same study failed to confirm a similar effect when CDEIS or global evaluation of lesion severity [GELS] was considered.[26]

One further issue, highlighted by our results, is the slightly lower inter-observer agreement of the Mayo endoscopic subscore, when compared with the Rutgeerts score or the SES-CD or CDEIS. This

Table 1. Individual values [for yearly data] and summary results [meta-analytical] with kappa statistics and 95% confidence interval [95% CI] for Mayo endoscopic subscore [Mayo] and Rutgeerts score [Rutgeerts] pre- and post-training programme, as well as intraclass correlation coefficient [ICC] with 95% CI for individual values [for yearly data] and summary results for Simple Endoscopic Score for Crohn's Disease [SES-CD] and Crohn's Disease Endoscopic Index of Severity [CDEIS].

	Year 1 results	Year 2 results	Overall results	*P* value
	121 observers, 5 videos	64 observers, 6 videos	185 observers	
Mayo pre-training	0.46 [0.41–0.50]	0.57 [0.51–0.62]	0.51 [0.48–0.55]	$P < 0.0001$
Mayo post-training	0.74 [0.70–0.77]	0.78 [0.72–0.83]	0.76 [0.72–0.79]	
	74 observers, 5 videos	69 observers, 6 videos	143 observers	
Rutgeerts pre-training	0.35 [0.27–0.42]	0.55 [0.49–0.61]	0.45 [0.40–0.50]	$P < 0.0001$
Rutgeerts post-training	0.85 [0.80–0.91]	0.72 [0.65–0.78]	0.79 [0.74–0.83]	
	137 observers, 5 videos	128 observers, 6 videos	265 observers	
SES-CD [single observation]	0.69 [0.44–0.95]	0.86 [0.69–0.98]	0.77 [0.56–0.96]	NA
	137 observers, 5 videos	100 observers, 6 videos	237 observers	
CDEIS [single observation]	0.70 [0.46–0.95]	0.84 [0.65–0.98]	0.76 [0.54–0.96]	NA

NA: not aplicable

issue is well recognized in the literature, and seems to be overcome when expert observers are involved.[1,9] It is equally patent that even if experts are considered,[10,11] kappa values for inter-observer agreement could be as low as 0.47–0.50. Results after the educational project reported in this study compare favourably with those previously mentioned, with kappa values as high as 0.76 [95% CI 0.72–0.79]. The higher discrepancies in ulcerative colitis scoring may find an explanation in the intrinsically wider variability of endoscopic pictures of ulcerative colitis, which may lead to a more complex and unreliable classification of disease severity.

The main conclusion of our study was was that there was a significant increase in inter-observer agreement thanks to a dedicated training process for two widely used IBD endoscopic scores [the Mayo endoscopic subscore and Rutgeerts score] and to peer discussion of the determinants and the various views on scoring a set of endoscopic videos. Our results suggest a possible alternative to diffuse central reading for increasing quality and reliability of endoscopic scoring in clinical practice and trials: education of regional endoscopists in a number of scoring conventions, and this may finally result in reducing disagreement.

Future studies supporting the effects of learning programmes on score agreement, involving larger batches of videos [with the goal of covering a larger moiety of video complexity] and with longer follow-up are warranted. Scientific societies in the future may do well to develop and maintain continuing medical education programs focused on supporting their members in reaching and maintaining good proficiency in IBD endoscopic scoring systems.

Funding

The IGIBDEndo educational project was funded through an unrestricted educational grant from AbbVie Italy.

Conflict of Interest

All the authors state they do not have any specific conflict of interest to disclose related to the present manuscript.

Acknowledgments

The Authors thank AbbVie Italy, who supported the Italian Group for Inflammatory Bowel Disease [IGIBD] through an unrestricted educational grant dedicated to funding the IGIBDEndo educational project.

Author Contributions

Study concept and design: Marco Daperno, Michele Comberlato, Fabrizio Bossa, Renata D'Incà, Ambrogio Orlando, and Fernando Rizzello. Acquisition of data: Marco Daperno, Michele Comberlato, Fabrizio Bossa, Armuzzi Alessandro, Livia Biancone, Andrea G Bonanomi, Rocco Cosintino, Giovanni Lombardi, Roberto Mangiarotti, Alfredo Papa, Roberta Pica, Renata D'Incà, Ambrogio Orlando, and Fernando Rizzello. Statistical analysis: Grassano Luca, and Guido Pagana. Analysis and interpretation of data: Marco Daperno, Grassano Luca, and Guido Pagana. Drafting of the manuscript: Marco Daperno, Michele Comberlato, Fabrizio Bossa, Grassano Luca, Guido Pagana, Renata D'Incà, Ambrogio Orlando, and Fernando Rizzello. Critical revision of the manuscript: Marco Daperno, Michele Comberlato, Fabrizio Bossa, Armuzzi Alessandro, Livia Biancone, Andrea G Bonanomi, Rocco Cosintino, Giovanni Lombardi, Roberto Mangiarotti, Alfredo Papa, Roberta Pica, Grassano Luca, Guido Pagana, Renata D'Incà, Ambrogio Orlando, and Fernando Rizzello.

References

1. 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 e2.
2. Rutgeerts P, Reinisch W, Colombel JF, *et al*. Agreement of site and central readings of ileocolonoscopic scores in Crohn's disease: comparison using data from the EXTEND trial. *Gastrointest Endosc* 2016;**83**:188–197 e3.
3. Baert F, Moortgat L, Van Assche G, *et al*. Mucosal healing predicts sustained clinical remission in patients with early-stage Crohn's disease. *Gastroenterology* 2010;**138**:463–8; quiz e10-1.
4. Colombel JF, Sandborn WJ, Reinisch W, *et al*. Infliximab, azathioprine, or combination therapy for Crohn's disease. *N Engl J Med* 2010;**362**:1383–95.
5. Manginot C, Baumann C, Peyrin-Biroulet L. An endoscopic Mayo score of 0 is associated with a lower risk of colectomy than a score of 1 in ulcerative colitis. *Gut* 2015;**64**:1181–2.
6. 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.
7. Peyrin-Biroulet L, Reinisch W, Colombel JF, *et al*. Clinical disease activity, C-reactive protein normalisation and mucosal healing in Crohn's disease in the SONIC trial. *Gut* 2014;**63**:88–95.
8. Levesque BG, Sandborn WJ, Ruel J, *et al*. Converging goals of treatment of inflammatory bowel disease from clinical trials and practice. *Gastroenterology* 2015;**148**:37–51 e1.
9. Osada T, Ohkusa T, Yokoyama T, *et al*. Comparison of several activity indices for the evaluation of endoscopic activity in UC: inter- and intraobserver consistency. *Inflamm Bowel Dis* 2010;**16**:192–7.

10. Travis SP, Schnell D, Feagan BG, *et al*. The impact of clinical information on the assessment of endoscopic activity: characteristics of the Ulcerative Colitis Endoscopic Index Of Severity [UCEIS]. *J Crohns Colitis* 2015;**9**:607–16.
11. Travis SP, Schnell D, Krzeski P, *et al*. Reliability and initial validation of the ulcerative colitis endoscopic index of severity. *Gastroenterology* 2013;**145**:987–95.
12. Khanna R, Zou G, D'Haens G, *et al*. Agreement among central readers in the evaluation of endoscopic disease activity in Crohn's disease. *Gut* 2016;**65**:1119–25.
13. Daperno M, Comberlato M, Bossa F, *et al*. Inter-observer agreement in endoscopic scoring systems: preliminary report of an ongoing study from the Italian Group for Inflammatory Bowel Disease (IG-IBD). *Dig Liver Dis* 2014;**46**:969–73.
14. Gecse KB, Löwenberg M, Bossuyt P, *et al*. Agreement among experts in the endoscopic evaluation of postoperative recurrence in Crohn's disease using the Rutgeerts score. *J Crohns Colitis* 2014;**8**:P285.
15. 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**:6.
16. 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.
17. Rutgeerts P, Geboes K, Vantrappen G, Beyls J, Kerremans R, Hiele M. Predictability of the postoperative course of Crohn's disease. *Gastroenterology* 1990;**99**:956–63.
18. 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 Therapeutiques des Affections Inflammatoires du Tube Digestif [GETAID). *Gut* 1989;**30**:983–9.
19. 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.
20. Cao H, Sen PK, Peery AF, Dellon ES. Assessing agreement with multiple raters on correlated kappa statistics. *Biom J* 2016;**58**:935–43.
21. Dignass A, Van Assche G, Lindsay JO, *et al*. The second European evidence-based Consensus on the diagnosis and management of Crohn's disease: current management. *J Crohns Colitis* 2010;**4**:28–62.
22. De Cruz P, Kamm MA, Hamilton AL, *et al*. Efficacy of thiopurines and adalimumab in preventing Crohn's disease recurrence in high-risk patients – a POCER study analysis. *Aliment Pharmacol Ther* 2015;**42**:867–79.
23. Orlando A, Mocciaro F, Renna S, *et al*. Early post-operative endoscopic recurrence in Crohn's disease patients: data from an Italian Group for the study of inflammatory bowel disease (IG-IBD) study on a large prospective multicenter cohort. *J Crohns Colitis* 2014;**8**:1217–21.
24. Reinisch W, Angelberger S, Petritsch W, *et al*. Azathioprine versus mesalazine for prevention of postoperative clinical recurrence in patients with Crohn's disease with endoscopic recurrence: efficacy and safety results of a randomised, double-blind, double-dummy, multicentre trial. *Gut* 2010;**59**:752–9.
25. Marteau P, Laharie D, Colombel JF, *et al*. Interobserver variation study of the Rutgeerts Score to assess endoscopic recurrence after surgery for Crohn's disease. *J Crohns Colitis* 2016;**10**:1001–5.
26. Dubcenco E, Zou G, Stitt L, *et al*. Effect of standardised scoring conventions on inter-rater reliability in the endoscopic evaluation of Crohn's disease. *J Crohns Colitis* 2016;**10**:1006–14.