Online Supplemental Material

To

Endoscopy and Central Reading in Inflammatory Bowel Disease Clinical Trials: Achievements, Challenges and Future Developments

1. Read Algorithms

1.1. Definition

Read algorithms are different from the endoscopic scoring system. They formalize 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.

When researchers apply a scoring system to clinical imaging, they implicitly assume an underlying construct that cannot be measured directly. In inflammatory bowel disease the underlying construct could be "disease severity," "inflammatory burden," etc. Regardless whether this is stated explicitly, the scoring system identifies criteria by which this construct is to be *measured*. Some scores may describe the construct more fully than others, but we leave this outside of scope for this discussion. We are here concerned about the final path, how the score of more than one reader for the same read instance can be integrated.

Measuring, observing and sampling are statistically closely related terms. However, "the strong temptation is, and always has been, to select one observation thought to be best, rather than to corrupt it by averaging it with others of suspected lesser value" (1). To clarify, the application of the "best observation" principle is analogous to judgment (adjudication), and, in contrast, averaging or other statistical data aggregation, assuming that all observations have been obtained by the same method, measurement.

While rheumatoid arthritis clinical trials, as one example, have subscribed to the *measurement* concept for years (two reader average for radiological progression) (2), the development of the equivalent read algorithm has been more circuitous in gastroenterology, specifically in inflammatory bowel disease trials, where providing for better *judgment* was the first attempt at improving the accuracy of central reads. Uncritical extrapolation of what works in one therapeutic area, however, to another must be avoided.

For example, consider the differences between central reading in rheumatoid arthritis and inflammatory bowel diseases (Table 1)

Table 1 Endpoints in RA and IBD clinical trials

	IBD	RA
Principal item of interest	Ulcerations	erosions, joint space narrowing
Used for eligibility	Frequently	Variable
One-directionality	May improve, worsen or fluctuate	Yes, always worse unless progression halted. Negative progression scores are considered erroneous.
Discrete anatomic locations	No, lesions can be anywhere and may, but not necessarily recur in the same location	Search space limited to predefined joint areas. Readers look at subtle changes in discrete locations.
Anatomic location exactly known and reproducible	Locations during colonoscopy can only be estimated	Yes
Patient preparation important	Yes, bowel prep	N/A
Analogy to well-known visual puzzles	"Find in picture"	"Spot the difference"

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 to 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.

1.2. Non-statistical (social) data aggregation

In a consensus-based approach a panel looks at the image or other matter of interest and comes, after open deliberation, to a conclusion. Without going into detail here, it can be agreed that 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 (3).

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, I.e., voting.

1.3. Statistical data aggregation: averaging and voting

Of the statistical data aggregation methods, two need to be discussed, averaging and voting. Here, accuracy improvements are transparent. For averaging, they follow a square root law which holds that

Gottlieb K, *et al. Gut* 2020;0:1–9. doi: 10.1136/gutjnl-2020-320690

the standard error of the sample mean decreases with the square root of the number of samples (36). 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 (4). Still, statistical data aggregation can be done using voting. The accuracy improvements using voting can also be described mathematically with the Condorcet Jury Theorem (5). 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 (5) and Ahmad et al. (6).

1.4. Empirical validation of 2+1 voting algorithms in clinical trials: Importance of the site reader

To our knowledge, the FITZROY study of filgotinib for Crohn's disease was the first clinical trial published that used more than one reader, but the data aggregation was based on adjudication (non-statistical), not voting, and the site endoscopist was not involved in the efficacy assessment reads, only during the eligibility determination. For eligibility reads, a single central reader was assigned. Only reads which were discrepant between the endoscopic assessment of the colonoscopy site (local reader) and central reader (eligible vs not eligible) were allocated to an adjudicator, who did a separate scoring, masked to the original reads. The result of the adjudicator's read was used as the final determination of eligibility. For the efficacy reads, assessment was done by two independent central readers in a masked fashion. In the case of discrepant results, a third, independent read by an adjudicator was done and was used as the final determination of efficacy read (7) . FITZROY met the primary endpoint of clinical remission based on CDAI, but none of the endoscopic endpoints based on the SES-CD.

Another approach was chosen for central reading in the exploratory cohort of the etrolizumab BERGAMOT study. Voting, not adjudication was employed (8). 300 patients with moderate to severe CD were randomly assigned to two doses of etrolizumab (pooled for analysis) or placebo for a 14-week induction period. The endoscopy videos were read by a local reader (LR) and two central readers (CRs); all readers were blinded to other readers' assessments and patients' clinical information. The local readers were trained in the scoring algorithm. Various endoscopy read models were used to assess endoscopic improvement (defined as ≥50% reduction of Simple Endoscopic Score for Crohn's Disease [SES-CD] from baseline) at week 14. Inclusion of at least two central readers decreased placebo rates and increased effect sizes with etrolizumab for endoscopic improvement in CD. Importantly, involving the LR and CR with a third reader (CR2) in cases of discordance (the 2+1 model) in the central read model resulted in the highest rate of endoscopic improvement with etrolizumab over placebo. In contrast, in FITZROY, where CDAI remission was significantly different compared with placebo, but SES-CD endpoints were not, site endoscopists were excluded from reading and there was an adjudicator (with final say, no voting) in case of discrepancies between 2 central readers.

The results of the analysis of the BERGAMOT data lend empirical support to the use of a 2+1 voting algorithm because they confirm the theoretically predicted superiority of voting and statistical data aggregation n general (as opposed to adjudication), and suggest that it is indeed critical to involve the site reader in the central read model. Site readers that know they are subject to supervision may not exhibit as much bias as original suggested by Feagan (9).

As the endoscopic component of the Mayo score for ulcerative colitis is on an ordinal scale, the voting approach may exist in a pure form. Other scores, for example the SES-CD above, approach a continuous

scale, and here 2+ 1 has been successfully modified by using a concordance threshold, i.e., considering a 2-point difference between two scores as the same 'vote', invoking a third read only when this allowable difference is exceeded (8).

2. Operational and logistics aspects of central reading

2.1. Temporal read sequence – paired, in sequence, or random

Images in several disease areas could be read in pairs, not knowing which one was acquired before or after the intervention, in sequence, or one at a time, i.e., random. The EMA guideline for rheumatoid arthritis clinical trials recommends that, in addition to being blinded regarding treatment allocation, "ideally, the readers should also … not be aware of baseline/previous images of the same subjects when scoring [a] later one" (10). Whether this is important, is subject to debate (11), (12), (13). Preliminary data generated by experienced central readers in a small subset of patients in the exploratory BERGAMOT induction cohort (etrolizumab in CD) showed that both paired and unpaired central reading of CD endoscopic videos were associated with very low placebo response rates. However, unpaired central reading was associated with increased treatment response rates compared to paired reads (14), Usually in IBD trials the baseline colonoscopy is read as soon as possible for enrollment qualification purposes, and not again, and the ready-pool of readers may have changed when the final colonoscopy has been obtained after 52 weeks. Enrollment or qualification and baseline assessment for the efficacy assessment can be disentangled, as we will discuss below. This may have the advantage that the same reader could read the same patient every time as there is no reader attrition if all videos are read in one campaign.

2.2. Central Reader Assignment

In the past, read companies have endeavored to assign the same IBD patient throughout the study to the same reader "as much as possible". The reason advanced for assigning the same reader to the same patient is that intra-reader variability is lower than inter-reader variability. However, while this sound good, in practice this may create more problems than it solves.

The goal of trying to have one reader read the same patient in large IBD clinical trials is in our experience in practice not accomplished 30 % of the time. Part of the reason for this is operational exigencies, e.g., the need for speed during eligibility and re-randomization determination. Also, during a one-year trial some 'early' readers may no longer be available. All of this introduces biases that are out of our control.

There are two solutions to this problem: First, insist in a complete random assignment for every reader-video match, or, alternatively, discard the eligibility central read for the efficacy assessment, if the readers would have to be different. In that case the efficacy read should be done with a new reader who can read all videos (the week 0 and efficacy time point ET1, ET2, etc.). This is done in rheumatoid arthritis trials and known as read campaigns. There are no compelling reasons why this should not be possible in IBD trials. Waiting until all efficacy assessments can be read together would ensure that one reader can always read baseline and post-intervention videos, even if they are 52 weeks apart. If eligibility determination and efficacy assessment are deconvoluted or separated, as we discuss here, still only one colonoscopy is needed, at time T=0. However, the eligibility reading could be done in any manner that is sound and operationally fast, even with a new score, or in a manner not yet developed, for example, using a machine learning algorithm. The possibility that there later may be discrepancies between the eligibility score and the efficacy score on the same video at T=0 deserves is of less

Gottlieb K, *et al. Gut* 2020;0:1–9. doi: 10.1136/gutjnl-2020-320690

importance, than sometimes feared. Even FDA does not elevate these topics to the level worthy of inclusion in their draft UC guidance: "Trial design issues such as the assessment of disease severity (on entry) … are beyond the scope of this guidance" (15) and the EMA advises that "… patients can be classified as having mild, moderate or severe disease activity according to one or more measures of disease severity" (16).

It is foundational to the 2+1 read model that readers other than the site reader are chosen at random, in an attempt to randomly distribute central reader biases across read cases. While we hold that complete randomization of central readers to cases indispensable, central read companies say that this is hard to do. Perhaps this is a result of having too small pool of central readers and a GI society directed broader central reader education program would be helpful in expanding that pool.

To ensure that random reader assignment is accomplished, every clinical trial report should include a demographics and characteristics table for the readers, analogous to a baseline characteristics and demographics table for patients. This table could include information how the reader was assigned across baseline, induction and maintenance reads and active and placebo to check for imbalances.

2.3. Optimal number of central readers per study, reader fatigue and rushing

One question which has not been examined is whether there is an optimal number of central readers per study and how long a reading session should be. There is evidence that batch reading improves performance in radiology, up to a point (17), but how large should batches be, and when is it time for a break? If a single reader can handle a total of 250 videos, as we surmise above, clearly, they cannot be read all at once. In the radiology literature "there is clear evidence that fatigue, particularly towards the end of a long shift, contributes to serious medical errors, and increases the risk of missing abnormalities on imaging" (17). In an analysis of 86,624 colonoscopies performed by 131 physicians at two medical centers, the adenoma detection rate for colonoscopies performed at the 9th + position was significantly lower than those at the 1st–4th or 5th–8th position, 27.2 (CI 25.8–28.6) versus 29.9 (CI 29.5–30.3), 30.2 (CI 29.6–30.9), respectively. Withdrawal time steadily decreased by colonoscopy position going from 11.6 (CI 11.4–11.9) min for the 1st colonoscopy to 9.6 (8.9–10.3) min for the 9th colonoscopy (18). The decreasing withdrawal time can be either interpreted as an effect of fatigue, or a conscious attempt to stay on schedule ('rushing'), or a combination of both. It would be naive to assume that these are issues that have no bearing on central reading. We also hear from central readers that sometimes videos are reviewed in settings where there are many distractions, such as airports, meetings, etc.

References

1. Stigler SM. The seven pillars of statistical wisdom. 2016.

2. Navarro-Compán V, Landewé R, Ahmad HA, Miller CG, Xu D, Wolterbeek R, et al. Rate of adjudication of radiological progression in rheumatoid arthritis randomized controlled trials depending on preset limits of agreement: a pooled analysis from 15 randomized trials. Rheumatology. 2013 Aug 1;52(8):1404–7.

3. Gottlieb K, Travis S, Feagan B, Hussain F, Sandborn WJ, Rutgeerts P. Central Reading of Endoscopy Endpoints in Inflammatory Bowel Disease Trials: Inflamm Bowel Dis. 2015 Jun;1.

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

5. 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 [Internet]. 2015 Feb 19 [cited 2018 Oct 28];15. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4349725/

6. 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 Feb 1;4(1):35–8.

7. Vermeire S, Schreiber S, Petryka R, Kuehbacher T, Hebuterne X, Roblin X, et al. Clinical remission in patients with moderate-to-severe Crohn's disease treated with filgotinib (the FITZROY study): results from a phase 2, double-blind, randomised, placebo-controlled trial. The Lancet. 2017 Jan;389(10066):266–75.

8. Reinisch W, Mishkin DS, Oh YS, Schreiber S, Maciuca R, Jacob R, et al. P132 Analysis of various central endoscopy reading methodologies in the BERGAMOT exploratory induction cohort evaluating etrolizumab in Crohn's Disease. J Crohns Colitis. 2018 Jan 16;12(supplement_1):S161–S161.

9. Feagan BG, Sandborn WJ, D'Haens G, Pola S, McDonald JWD, Rutgeerts P, et al. The Role of Centralized Reading of Endoscopy in a Randomized Controlled Trial of Mesalamine for Ulcerative Colitis. Gastroenterology. 2013 Jul 1;145(1):149-157.e2.

10. European Medicines Agency. Clinical Investigation of Medicinal Products for the Treatment of Rheumatoid Arthritis [Internet]. 2018 Jan p. 16. Report No.: CPMP/EWP/556/95 Rev. 2. Available from: https://www.ema.europa.eu/en/clinical-investigation-medicinal-products-treatment-rheumatoid-arthritis

11. Salaffi F, Carotti M. Interobserver variation in quantitative analysis of hand radiographs in rheumatoid arthritis: comparison of 3 different reading procedures. J Rheumatol. 1997 Oct;24(10):2055–6.

12. Ferrara R, Priolo F, Cammisa M, Bacarini L, Cerase A, Pasero G, et al. Clinical trials in rheumatoid arthritis: methodological suggestions for assessing radiographs arising from the GRISAR study. Ann Rheum Dis. 1997 Oct 1;56(10):608–12.

13. van der Heijde D, Boonen A, Boers M, Kostense P, van der Linden S. Reading radiographs in chronological order, in pairs or as single films has important implications for the discriminative power of rheumatoid arthritis clinical trials. Rheumatology. 1999 Dec 1;38(12):1213–20.

14. Mishkin D, Schreiber S, Daperno M, Oh Y, Hussain F, Hassanali A, et al. Evaluation of Paired and Separated Endoscopic Reads at Baseline and Week 14 in the Exploratory BERGAMOT Induction Cohort: 856. Am J Gastroenterol. 2019;114.

15. Food and Drug Administration. Ulcerative Colitis: Clinical Trial Endpoints Guidance for Industry. Draft Guidance. [Internet]. Food and Drug Administration; 2016. Available from: https://www.fda.gov/media/99526/download

16. European Medicines Agency. Guideline on the development of new medicinal products for the treatment of Ulcerative Colitis. 2018 Jun 28;18.

17. Taylor-Phillips S, Stinton C. Fatigue in radiology: a fertile area for future research. Br J Radiol. 2019 Jul;92(1099):20190043.

18. Marcondes FO, Gourevitch RA, Schoen RE, Crockett SD, Morris M, Mehrotra A. Adenoma Detection Rate Falls at the End of the Day in a Large Multi-site Sample. Dig Dis Sci. 2018 Apr 1;63(4):856–9.