OPEN

Central Reading of Endoscopy Endpoints in Inflammatory Bowel Disease Trials

Klaus Gottlieb, MD, Simon Travis, MD,† Brian Feagan, MD,‡ Fez Hussain, MD,* William J. Sandborn, MD,§ and Paul Rutgeerts, MD, PhD||*

Background: Central reading of endoscopy (CROE) is crucial in determining who qualifies for a trial but also has a role, independent of the selected scoring system, in decreasing measurement noise that can obscure separation between placebo and active drug. Benefits of CROE may not be independent of the method chosen, and controversy exists about the ideal approach.

Methods: Literature review and concept development.

Results: Components to be considered in the reading algorithm are blinding, number of central readers, independent voting versus consensus panel, video recordings versus still images, and involvement of the site reader. Key concepts considered are endpoints, bias, power, and sample size derived from the Food and Drug Administration and European Medicines Agency guidelines, as well as the technological requirements and recruitment, qualification, and revalidation of central readers as applied to CROE.

Conclusions: Recording and CROE should be standardized, and an imaging charter developed with research on the different components and its overall impact.

(Inflamm Bowel Dis 2015;21:2475–2482)

Key Words: endoscopy, central reading, ulcerative colitis, Crohn's disease, clinical trials, endpoints

Clinical trial endoscopy and the central reading of endoscopy (CROE) are emerging fields, which offer much more opportunity for research but should also be of interest to the practicing gastroenterologist. More endoscopists will be needed to participate in clinical trials, either directly as investigators or as central readers, and instruments or indices used in clinical trials can readily be incorporated into clinical practice. Nevertheless, there are many unresolved issues concerning CROE that need to be addressed by investigators, regulatory authorities, the pharmaceutical industry, and clinical research organizations. In this review, we consider key components, evolving concepts, technological requirements and the recruitment, qualification, and revalidation of central readers. It is clear that there should be standardization of the central reading process, so a charter for CROE is proposed. This in turn will allow research into what works, what matters, and novel approaches.

Key Facts About Registration Trials

Clinical trials are customarily divided into phases I, II, and III. Phase I evaluates safety, so subjects are often healthy volunteers; some phase I studies are performed on patients with inflammatory bowel disease (IBD), but CROE matters little beyond the need to select a responsive endoscopic index and record videos because very few investigators are involved. Phase II studies evaluate different doses for efficacy in actual patients, and there are many design options not only a responsive endoscopic index needed but also a central reading potentially reduces variance between observers and enhances the signal to noise ratio. Phase III trials are "confirmatory" to apply for marketing authorization, and central reading is now frequently performed to reduce variance or inappropriate recruitment. In phase III, regulatory agencies promulgate disease-specific guidelines that define the relevant indications, conduct of the trials, the applicable patient population, inclusion and exclusion criteria, safety evaluations, and, most

Received for publication March 17, 2015; Accepted April 3, 2015.

From the *Gastroenterology Center of Excellence, Quintiles, Durham, North Carolina; †Translational Gastroenterology Unit, Oxford University Hospitals, Oxford, United Kingdom; ‡Robarts Research Institute, Western University, London, ON, Canada; §Division of Gastroenterology, University of California San Diego, San Diego, California; and ||Division of Translational Research in Gastrointestinal Disorders, University of Leuven, Leuven, Belgium.

K. Gottlieb and F. Hussain work for Quintiles, B. Feagan for Robarts Clinical Trials, companies that provide outsourcing services for the biopharmaceutical industry including central reading services or consulting. W. J. Sandborn provides consulting services to University of Western Ontario (owner of Robarts Clinical Trials). S. Travis and P. Rutgeerts provide central reader services and consulting to Quintiles and/or Robarts Clinical Trials.

Reprints: Klaus Gottlieb, MD, Synthetic Biologics, Inc., 617 Detroit Street, Suite 100, Ann Arbor, MI 48104 (e-mail: kgottlieb@syntheticbiologics.com).

DOI 10.1097/MIB.0000000000000470

Published online 18 June 2015.

importantly, the outcome measures. Phase IV studies occur after marketing to refine indications or to evaluate benefits or risks in specific patient populations. As with phase III studies, CROE should become the norm to confirm appropriate patient selection. Nevertheless, country-specific phase IV studies present a particular challenge to central reading not only because of the available pool of experienced central readers and the costs involved. Costs can be offset by the effect of central reading on enriching the patient population and reducing the placebo response (see below).

"Outcome measures" is an overarching term, which includes "clinically meaningful endpoints," "patient-reported outcomes," "biomarkers" and "surrogate endpoints." Clinically meaningful endpoints describe or relate to the way a patient feels, functions, and survives.[1] A surrogate endpoint is a marker that is believed to predict clinical benefit but is not itself a direct measure of clinical benefit, such as a decrease in tumor size.[2] Although endoscopy is a contender, there are currently no "surrogate endpoints" acceptable to regulatory agencies for trials in the most gastrointestinal diseases.

Endoscopic Endpoints

There is an emerging consensus in the regulatory and academic community that trials in general and IBD in particular, need endpoints that encompass both the patients' experience (as measured by a patient-reported outcome instrument, PRO) and an imaging endpoint (mucosal healing, measured or demonstrated by endoscopy) with or without histopathology.[3] These 2 factors could be achieved by a "coprimary endpoint" (where both endpoints need to be achieved simultaneously by each patient) or by an arrangement where the PRO is the primary endpoint, and the endoscopy endpoint becomes the "first-ranked secondary endpoint" (which is now acceptable to the European Medicines Agency, EMA).[4] Although the Food and Drug Administration (FDA) has yet to release updated guidelines on trial conduct for ulcerative colitis (UC) or Crohn's disease (CD), they are expected strongly to favor a coprimary endpoint, consisting of a PRO and an endoscopic scale for both diseases. Coprimary endpoints are more difficult to achieve but may be necessary when symptoms reported by patients do not reliably correlate with the underlying inflammation, as is the case with CD. The choice of trial endpoint influences sample size estimation for both noninferiority design and superiority trial designs. Adequate treatment of either trial design is outside the scope of this review, but the following points relate to endoscopic or other endpoints.

Power, Sample Size, and Central Reading

Power is the chance of detecting a departure from the null hypothesis at a given alpha (type I) error if there a difference truly exists between drug and placebo. Power is influenced by the choice of statistical test, the sample size, size of the experimental effect, and, most importantly for endoscopy, the level of error in the experimental assessment. Total error is composed of systematic directional error (bias) and random error. Control of these 2 sources of error in a central-read process may require different approaches. Size of the experimental effect or effect size is commonly and in Table 1 defined as the proportion (percentage) of patients who responded to active drug minus the proportion (percentage) of patients who responded to the placebo. Often, this is also called the separation between drug and placebo.

The desired power is set at a specific level, often 0.80, and the sample size to accomplish this power is then determined before the trial starts. The necessary power can be achieved with a smaller sample size by increasing the accuracy of the outcome measurement instead of increasing the sample size. In Table 1, we have calculated power retrospectively to illustrate one point: the choice of the measurement instrument, here the central reading process, can influence power. In phase IIb and III registration trials in IBD, endpoints are not measured on continuous scales. Instead, binary definitions of "responders" or "remitters" are recorded (defined, e.g., by the Mayo Clinic Score) as respective proportions for drug and placebo. Table 1, based on published data, illustrates that improving the accuracy of outcome measurements minimizes sample size,[5] sometimes this may, however, not be the case depending on the criteria used in the reading algorithm.[6] Central reading limits bias that site endoscopists may have especially when the endoscopist is

TABLE 1. Example How Central Reading Can Influence Effect Size and Sample Size

	Feagan et al[5] (Oral Mesalamine)		Kobayashi et al[6] (Rectal Mesalamine)	
	MCS Remission: Site Readers	MCS Remission: Central Readers	Endoscopic Remission: Site Readers	Endoscopic Remission: Central Readers
Drug	0.300	0.29	0.828	0.906
Placebo	0.206	0.138	0.311	0.590
"Effect" size	0.094	0.152	0.517	0.316
Retrospectively calculated sample size for power = 0.80 (both arms)[7]	670	228	28	58

MCS: mayo clinic score, which includes endoscopic assessment.

also the principal investigator and/or the personal physician of a patient who is a trial candidate. When evaluating Table 1, it should be remembered that studies with lower disease activity scores at study entry tend to have higher placebo remission rates and vice versa.[8] Both studies were conducted in a population with mild-to-moderate disease.

Table 1 shows that central reading in one trial (Feagan et al) reduced the observed placebo response rate from 20.6% (0.206, site readers) to 13.8% (0.138, central readers), which increased the difference from placebo to active drug (effect size) from 9.4% to 15.2%. This increased the power of the study and is associated with a lowered retrospectively calculated sample size (see Discussion). In contrast, in the other study (Kobayashi et al), the effect of central reading was the opposite: Separation between placebo and active drug (effect size) was reduced, which resulted in a decrease in study power and a higher retrospectively calculated sample size.

"Accuracy" and "effect of bias" are embedded in the separation ("spread") between the proportions of patients who respond to placebo versus drug. This spread is often called the "effect size." Effect size is therefore a composite of the actual efficacy of the drug and the performance characteristics of the instrument for outcome measurement. The example (Table 1) shows that in one study, central reading increases the effect size, whereas in the other study, the effect size decreases. The studies are not directly comparable because one used independent assessment of videotaped colonoscopies[5] and the other used still photographs taken at endoscopy and a consensus panel rather than readers who scored independently of each other.[6] In both cases, the intrinsic activity of the drug was obviously not changed by central reading, but something happened to bias and/or accuracy. This indicates that central reading by itself cannot be expected automatically to lead to increased study power or a decreased required sample size.

Noninferiority Trials

Most registration trials are controlled trials of superiority over placebo, with the aim of rejecting the null hypothesis that there is no difference between the two. If 2 active drugs are compared, trial designers have the option of aspiring to show superiority, equivalence, or noninferiority. These concepts are explored in a published FDA guidance document.[9] Two points are relevant to endoscopic endpoints: sample sizes required for noninferiority studies are much larger than those for superiority studies (and those for equivalence much larger still). The guidance states that in contrast to superiority trials, "in noninferiority trials, many kinds of problems fatal to a superiority trial, such as measurement problems more generally (i.e., "noise") can bias towards noninferiority where it did not really exist."[9] Consequently regulators are particularly interested in how this "noise" was limited. Reducing "noise" concerns the accuracy of measurements, which is where specific details of the central reading process (technology, reader expertise, details of the imaging charter, and quality controls) become paramount.

CENTRAL READING OF ENDOSCOPY

Independent, central, blinded review or reading of imaging endpoints in clinical trials dates back several decades. It was, for example, used in the National Cooperative Gallstone Study, where oral cholecystograms, performed locally according to a standardized protocol, were evaluated by a central radiology unit for eligibility and, subsequently, for efficacy.[10] Conceptually, blinded independent central review could be extended to any endpoint that has a subjective component with the obvious exception of PROs.[11] Whether this paradigm is always accompanied by efficiency, gains has been debated.[12] We will limit the appraisal to central reading of imaging endpoints.

In contrast to trials that use radiological endpoint assessments, CROE is a relatively new development. To our knowledge, CROE to monitor the consistency of endoscopic assessment by site investigators in UC was first reported as a meeting abstract by Abreu et al in 2006. The blinded off-site reader disagreed with investigator scoring for endoscopic disease severity at baseline by 12% to 23% across two-dose levels of investigational agent compared with mesalamine.[13] As a consequence, the case was made for standardization of reading procedures.[14] This abstract was followed in 2009 by evaluation of the results of a multicenter trial of delayed-release oral mesalamine. Although no assessment of the potential effects of central reading on the outcome of this noninferiority trial was performed,[15] awareness of variability in endoscopic assessment drove an international group to use the videos in this study to develop the first validated index of endoscopic activity in UC.[16,17] In 2013, Feagan et al[5] demonstrated how central reading could alter regulatory endpoints for licensing a new formulation of mesalamine, and in 2014, Kobayashi et al[6] described the impact of central reading of still photographs on a trial of topical mesalamine.

In trials of CD, central reading of video colonoscopy became established after the MUSIC trial evaluated therapy with certolizumab pegol for treatment of active disease. The MUSIC study suggested that centrally read scores for the Crohn's Disease Endoscopic Index of Severity were lower compared with local readings, particularly at baseline.[18] Although the differences between central readers and site readers were not statistically significant, they were numerically lower at 3 time points. Subsequently, the EXTEND study evaluating adalimumab therapy for active CD became the first study in CD to use an endoscopic primary endpoint and ensured that the Crohn's Disease Endoscopic Index of Severity was determined by a blinded central reviewer using videos obtained by site endoscopists[19] who were not involved in the reading process.

TERMINOLOGY

Central reading is often understood to mean that the interpretation of imaging is not or not only done by one individual, the site reader (endoscopist at the clinical trial site), but instead is supervised, amended, or adjudicated by at least 1 off-site other reader. The "central reader" is expected to have less bias or more expertise than the site reader. Blinding of the central reader and no

direct patient contact is considered to be indispensable and is now required by regulatory agencies. For example, as suggested by Feagan et al[5] and Hébuterne,[18] site readers may be biased towards upcoding the reads to allow their patients to meet entry criteria for severity in clinical trials. Nevertheless, the term "central" does not reflect current implementation models but will probably continue to be used. The FDA in one of their guidances appears to prefer the term "off-site reading" and defines it as follows: "… off-site image evaluations are image evaluations performed at sites that have not otherwise been involved in the conduct of the study and by readers who have not had contact with patients, investigators, or other individuals involved in the study."[20]

Placebo Response Rates and Benefits of Central Reading

Average placebo response rates in clinical trials of UC were reported in a meta-analysis published in 2007 as 13% (95% confidence interval [CI], 9%–18%) for clinical remission and 28% (95% CI, 23%–33%) for response.[8] Both endpoints are based on the Mayo Clinic Score.[20] An updated meta-analysis that included trials up to 2014 reported figures of 10% (95% CI, 7%–13%) for remission and 33% (95% CI, 28%–38%). Approaches to reduce placebo rates include enrichment strategies based on endoscopic activity thresholds, exclusion of patients with no histological disease activity, or CRP thresholds to select patients who have objective evidence of inflammation at baseline rather than symptoms alone.[21] The most recently completed UC registration trial (vedolizumab) showed a placebo response rates of 5.4% for clinical remission and 25.5% for clinical response.[22] The vedolizumab trials were conducted without central reading. The etrolizumab phase 2 trial seems to have been the first prospective, randomized, placebo-controlled trial in which a centrally read Mayo Clinic Score endoscopy subscore of at least 2 was an eligibility requirement. Clinical remission and response were measured both at weeks 6 and 10. For clinical remission, placebo rates were 5% and 0% at weeks 6 and 10, respectively, and for clinical response, respective rates were 34% and 29%.[23] If central reading can help reliably restrict placebo remission rates to low single figures (0%–5%), then the need for a placebo comparator in such trials might be questioned and active comparators sought.

Table 1 shows that an increased effect size (drug response–placebo response rate) reduces the required sample size needed for a given power or from a different perspective achieves greater statistical power. Table 2 shows the sample sizes needed for different combinations of placebo and drug response rates. Readers will get a sense of the nonlinear relationships and will appreciate that modest increases in the effect size can result in significant reductions of the sample size.

Although there are good theoretical reasons to expect that central reading improves accuracy when both more than 1 reader and a voting scheme are used,[25] it should not be assumed that central reading by itself will necessarily increase the effect size by lowering placebo response rates. That is because the reading scheme ("algorithm") and other implementation details (e.g., a single central reader, image type and quality, adjudication of difference through a voting system or consensus committee) could have a material impact. We will return to this in the section on the research agenda.

REGULATORY ENVIRONMENT

Food and Drug Administration

There is currently no published FDA guidance relating to CROE although it was broadly endorsed by the academic participants of the GREAT II conference in September 2012,

TABLE 2. Sample Sizes to Detect a Difference in 2 Proportions at a 5% Significance Level with 80% Power

Vertical: Placebo Response Rates	Horizontal: Drug Response Rates																			
	0.05	0.10	0.15	0.2	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
0.00	152	74	48	35	27	22	18	15	13	11	10	8	7	6	6	5	4	4	3	2
0.05		435	141	76	49	36	27	22	18	15	12	11	9	8	7	6	5	4	4	3
0.10			686	199	100	62	43	32	25	20	16	14	11	10	8	7	6	5	4	4
0.15				906	250	121	73	49	36	27	22	17	14	12	10	8	7	6	5	4
0.20					1094	294	138	82	54	39	29	23	18	15	12	10	8	7	6	5
0.25						1251	329	152	89	58	41	31	24	19	15	12	10	8	7	6
0.30							1377	356	163	93	61	42	31	24	19	15	12	10	8	6
0.35								1471	376	170	96	62	43	31	24	18	14	11	9	7
0.40									1534	388	173	97	62	42	31	23	17	14	11	8
0.45										1565	392	173	96	61	41	29	22	16	12	10

Adapted from Ref. 24. Adaptations are themselves works protected by copyright. So in order to publish this adaptation, authorization must be obtained both from the owner of the copyright in the original work and from the owner of copyright in the translation or adaptation.

hosted by the FDA Division of Gastroenterology Products.[3] Interestingly, a few months earlier in July 2012, an Oncology Drug Advisory Committee had concluded that the need for central radiological review of all subjects in studies that rely on progression-free survival as an endpoint had not clearly been demonstrated and that "audits of a percentage of the cases" might be sufficient.[26–28] Indeed, Dodd et al[29] have argued that blinded independent central review for progression-free survival removes some biases while introducing others. These issues may, however, be specific to oncology trials, which are often single arm and rely on survival statistics. These differ from placebo-controlled trials in IBD whose primary endpoints compare proportions of responders at predetermined time points. CROE is currently supported by expert opinion on theoretical grounds and by limited practical experience in IBD clinical trials. Additional work is needed to determine how "read paradigms" can be optimized to realize the benefits of CROE. Attempts by regulatory authorities to define these processes may yet be premature. A dialog among interested parties is needed.

Three guidance documents cover central reading in other specialties or drug development areas (Table 3). Many of these concepts might be adopted for CROE. Nevertheless, their relevance may be limited because endoscopic "imaging" is fundamentally different from radiological imaging that is acquired by a machine and not normally read in "real-time." Furthermore, video imaging that are obtained by technicians after a highly standardized script (e.g., echocardiography or transabdominal ultrasound) are also not comparable with endoscopy videos. Although this may appear obvious to the readers of this journal, it may not be so to nonendoscopists.

TABLE 3. FDA Guidances that Pertain to Blinded Independent Central Review Outside of Gastroenterology

Guidance	Selected Key Message
Guidance for Industry: Standards for Clinical Trial Imaging Endpoints (Draft Guidance) (2011)[30]	The need for a centralized (core) image interpretation process is contingent upon the role of imaging within the trial. In situations where image interpretation results in measurements representing important components of trial eligibility determination or safety or efficacy endpoints and these measurements are vulnerable to considerable variability among clinical sites, a centralized image interpretation process is needed
Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics (2007)[31]	Centralized independent verification of tumor endpoint assessments (especially for Progression- or Disease-Free Survival) may not be necessary when randomized trials are blinded or effect sizes are robust in large randomized trials where sensitivity analysis supports lack of observer bias
Guidance for Industry: Developing Medical Imaging Drug and Biological Products Part 3: Design, Analysis, and Interpretation of Clinical Studies (2004)[32]	Image interpretation is inherently subjective. Therefore, interreader variability and the need for adjudication are expected. The same images might be interpreted differently by central as opposed to local readers at a clinical site. We recommend the use of quantitative measurement of reader variability as a valuable index of reader performance

European Medicines Agency

Like the FDA, EMA has no guidance relating specifically to central reading in endoscopy. The definition of off-site reading is virtually identical to that used by the FDA. In addition, EMA recommends that off-site assessment should be performed by "a representative sample (2 or more) of readers." Sample means in this context that both readers evaluate the imaging independently not in conference. This is in contradistinction to consensus reads, which may be performed after individual readings have been completed. Readers are not independent anymore; consensus reads should not serve as primary image evaluation.[33] Such guidance raises important questions about CROE paradigms and adjudication in addition to trying to resolve differences between the FDA and EMA.

READING PARADIGMS

Central reading can be implemented in many different ways,[34] and any discussion about the possible merit of CROE is incomplete without close attention to how it is implemented. The articles by Feagan et al and Kobayashi et al are good examples of 2 very different approaches. In the first case, a single central reader was involved "for quality control and site training" and compared post hoc with the performance of the site readers.[5] In the second case, a central review committee was convened that consisted of 7 gastroenterologists in the field of IBD who were not participating in the trial ("central review members"). Each member reviewed the mucosal findings on 4 colonoscopic digital images (still photographies) submitted by the attending physicians. The results of their evaluations were then summed, and when at least 5 members agreed on a score for a case ("agreed group"), this score was adopted as the committee's review."[6] In cases where agreement was not reached, an adjudication panel was convened that consisted of at least 5 members of the 7-member team above, and the images were discussed until consensus was reached.[6]

Classification of CROE-read Paradigms

There are many options, so we propose the following classification of CROE schemes, which should cover most

practical applications. Eligibility determination is the most important component (Fig. 1).

Consensus Panels

To limit bias, EMA advises that consensus panels, as used by Kobayashi et al,[6] should not be used for primary image evaluation. Because this group also used still images instead of video recordings, this approach puts the central readers at a distinct disadvantage compared with the site reader. It is such factors that are likely to have contributed to the loss of separation between placebo and active drug as a consequence of central reading. Although a consensus panel that discusses imaging findings and their interpretation may indeed agree on a specific read, perhaps even unanimously, this consensus may be forced (i.e., biased) by the most vocal or powerful member of the panel. Central readers should come to their assessment uninfluenced by others. Although blinded central readers may not have a specific upcoding bias, it is likely that they may have other biases. This is why EMA advises that off-site assessment should be done by "a representative sample (2 or more) of readers"[33] rather than just 1 alone. This in turn appears to call for a system of voting, where each vote has equal weight and the majority determines the final score with a mechanism that can handle ties or other exceptions.[25]

Role of Site Readers

Site readers clearly have a primary role in initial patient evaluation, but they may also be involved in the central reading process. An argument against such involvement is that site readers may systematically upcode or downcode scores at determination. This applies both to the baseline read of an induction study and the primary endpoint of an induction trial, especially where the primary endpoint of the induction trial determines whether the patient is rerandomized into a blinded maintenance trial or moves into open-label therapy. Such upcoding or downcoding may occur consciously to allow more patients to enter the trial or perhaps unconsciously. Although evidence has been presented that this might be the case,[5,18] arrangements have been proposed to minimize this effect.[25]

Nevertheless, the site reader may yet be an important or even essential part of the reading algorithm.[25] The site endoscopist controls the quality of the video recording and lays the foundation for accurate reading for any subsequent evaluations. Better videos may be obtained by asking the site reader to record a video and score it with the knowledge that this will be confirmed or challenged by off-site readers. Proponents of this approach[25] state that it has 3 advantages:

1. Conscious upcoding or downcoding will decrease (i.e., scoring accuracy will increase).
2. The video recording quality will increase allowing a better performance by the off-site readers.
3. There is a training benefit to "watching over your shoulder."[35]

This read scheme is outlined in Figure 1. When site reader and first central reader agree, a final score can be determined. In cases of disagreement, a second central reader will be engaged who does not know the previous scores and is also unaware that this is the third read (2 + 1 algorithm). In the majority of cases, 2 scores out of 3 will be identical and a final overall score can be reported by majority voting. Details regarding the theoretical foundations for this algorithm and its implementation, including the handling of exceptions, were recently described by Gottlieb and Hussain.[25]

Technology Requirements

Technological implementations will continue to evolve in an unpredictable manner. Nevertheless, some trends in IBD trials are foreseeable. First, large-scale confirmatory phase 3 trials will continue to be conducted in multiple countries around

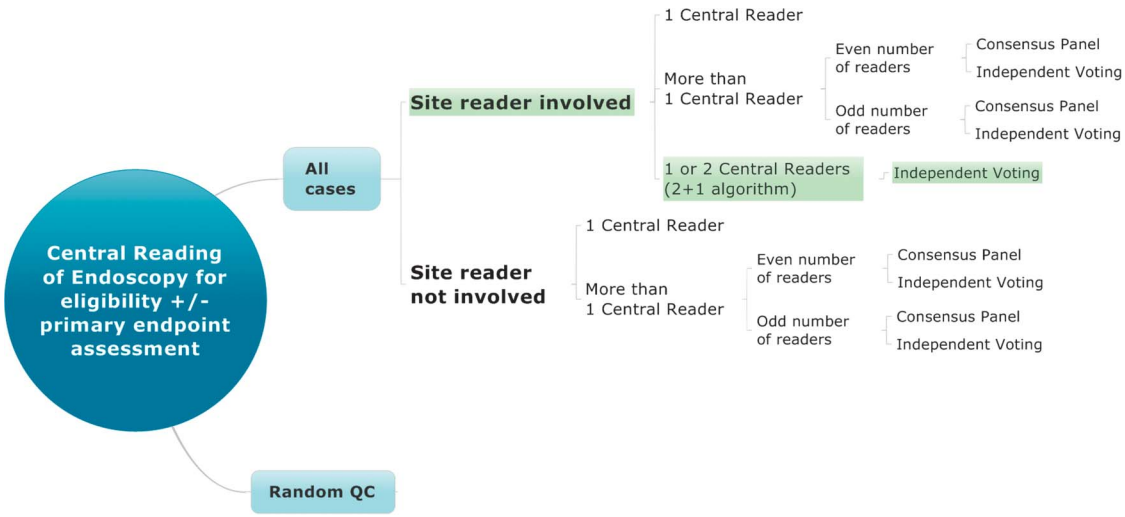

FIGURE 1. A general classification of central reading algorithms.

the world. Technology platforms will be different, but minimum capabilities need to be stipulated and be part of the site qualification process. Second, site endoscopes will typically also undergo a qualification procedure that consists of the recording of a test video that will be evaluated by the central reading provider (CRP), who will most often be a clinical research organization or one of their specialized vendors. Site readers need to record the videos using a standard protocol, transmit it to the CRP who then assigns the off-site (=central) readers. Third, the central readers need to adhere to the reading charter and have standardized viewing conditions. For more details see Ahmad et al and others in the peer-reviewed literature,[25,34,36] white papers[37] and FDA guidance documents.[32]

RECRUITMENT AND QUALIFICATIONS OF CENTRAL READERS

Although it remains to be established that all central reading algorithms decrease bias and noise, it is clear that there is a strong pressure for central reading. Consequently, many central readers will be required for numerous phase 2 and phase 3 trials. This begs several questions. How are endoscopists best trained in the selected scoring systems (Mayo Clinic Score, UCEIS, Crohn's Disease Endoscopic Index of Severity, etc.)? What should be the minimum baseline qualifications to become "central readers?" And what ongoing quality control measures are required? It has been argued that endoscopic scoring is a classification task with most disagreements occurring in boundary (borderline) situations. Although possibly true, this is not necessarily the case. It has also been argued that readers with considerable experience and well trained in the methodology do not perform differently than internationally recognized IBD specialists.[25] This is demonstrably untrue because agreement between such specialists can be poor depending on the instrument and related training. We believe that central readers should be practicing endoscopists with experience in the disease area in question who, after training, pass a standardized test and are periodically retested.

Standardization of Recording and Reading

Endoscopic scoring of mucosal disease depends greatly on the quality of the recordings. Standardization of the bowel preparation, duration of recording, recording on insertion or withdrawal of the scope, identification of segments, to name a few, are critical to a reliable and reproducible scoring of endoscopy recordings. Readers need to be instructed and agree on the scoring guidelines. This is more difficult for CD than UC because many issues require prior agreement, including how should scoring of lesions on the ileocecal valve or ileocolonic anastomosis be recorded (right colon or ileum?), what happens in cases of a long stenosis (to what segment will the score be assigned?), and others. Such considerations should be part of a trial imaging charter, which may require amending as unconsidered issues emerge during the trial.

IMAGING CHARTER

According to FDA guidance, "sponsors should generally develop a document that provides a comprehensive and detailed description of the clinical trial imaging methodology if a trial standard for image acquisition and interpretation applies to the imaging data. We suggest that sponsors refer to this document as an imaging charter and develop the document with the same care and attention to detail that is typically applied to the main components of a clinical protocol. Indeed, sponsors should generally regard the imaging charter as an integral component of the protocol, much as a statistical analysis plan is often developed as a component of a clinical protocol."[30] This guidance, although not geared towards endoscopy, still contains useful recommendation regarding the necessary components of such a charter.

SUMMARY AND RESEARCH QUESTIONS

The question "is central reading necessary" has largely been answered. However, greater clarity is needed regarding reading algorithms. We contrasted 2 examples, one where central reading was associated with a larger separation between drug and placebo, and another in which the opposite effect was observed. We believe that the disappointing results of the latter study were related to design features that we would not recommend, particularly the use of still images and consensus panels. Whether specific reading algorithms work or not is no small matter. There is a potential for substantially increased study power and significant cost savings if the algorithm works as intended. Conversely, added expense could not only fail to deliver any additional benefit but could in fact decrease the chances of a successful trial.

Clinical research organizations may be the most interested in determining which central reading algorithms work, because they are eager to keep clinical trials as lean as possible and pass cost savings on to their clients to stay competitive. In addition, the academic clinical trial community may be interested in pursuing such research, a wide open field, not least for early phase studies on drug development. Large clinical trials are underway where voting algorithms are used, and the performance characteristics of these algorithms need to be described in detail. The training and qualification process for central readers remains largely unexamined, but there seems little doubt that CROE is here to stay. Although endoscopic scoring systems are now an important part of both UC and CD clinical trials, their adoption in clinical practice remains low. This could be problematic if powerful drugs with their inherent risks are used in patients who were not studied in the clinical trials, i.e., do not meet the now prevailing (endoscopic) inclusion and exclusion criteria. However, there is some light on the horizon, and Daperno et al[38] have shown that community gastroenterologist are both willing and able to learn the respective systems through relatively limited interventions. More work in this direction is also needed.[39]

REFERENCES

1. Fleming TR, Powers JH. Biomarkers and surrogate endpoints in clinical trials. *Stat Med.* 2012;31:2973–2984.
2. Gottlieb K, Randazzo G, Walsh M. Chapter 11-Prescription Drug Product Submissions. Fundamentals of US Regulatory Affairs [Internet]. 8th ed. Regulatory Affairs Professionals Society; 2013:121–138. Available at: http://www.raps.org/WorkArea/DownloadAsset.aspx?id=5190. Accessed January 7, 2015.
3. Hyams JS. "GREAT"er than Last Year: GREAT 2 | AGA Washington Insider [Internet]. Available at: http://agapolicyblog.org/2013/11/18/greater-than-last-year-great-2/. Accessed January 7, 2015.
4. European Medicines Agency. Guideline on the Development of New Medicinal Products for the Treatment of Ulcerative Colitis [Internet]. European Medicines Agency; 2008. Report No.: CHMP/EWP/18463/2006. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003266.pdf. Accessed January 7, 2015.
5. 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;145:149–157.e2.
6. Kobayashi K, Hirai F, Naganuma M, et al. A randomized clinical trial of mesalazine suppository: the usefulness and problems of central review of evaluations of colonic mucosal findings. *J Crohns Colitis*. 2014. Available at: http://www.sciencedirect.com/science/article/pii/S1873994614001780. Accessed January 7, 2015.
7. Rosner B. *Hypothesis Testing: Categorical Data—Estimation of Sample Size and Power for Comparing Two Binomial Proportions. Fundamentals of Biostatistics*. Boston, MA: Brooks/Cole, Cengage Learning; 2011.
8. Su C, Lewis JD, Goldberg B, et al. A meta-analysis of the placebo rates of remission and response in clinical trials of active ulcerative colitis. *Gastroenterology*. 2007;132:516–526.
9. FDA. Guidance for Industry Non-inferiority Clinical Trials. Draft [Internet]. FDA; 2010. Available at: http://www.fda.gov/downloads/Drugs/Guidances/UCM202140.pdf. Accessed January 7, 2015.
10. Lachin JM, Marks JW, Schoenfield LJ, et al. Design and methodological considerations in the National Cooperative Gallstone Study: a multicenter clinical trial. *Control Clin Trials.* 1981;2:177–229.
11. Walovitch RC, Yao B, Chokron P, et al. Subjective endpoints in clinical trials: the case for blinded independent central review. *Open Access J Clin Trials.* 2013;5:111.
12. Hata J, Arima H, Zoungas S, et al. Effects of the endpoint adjudication process on the results of a randomised controlled trial: the ADVANCE trial. *PLoS One*. 2013;8:e55807.
13. 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. *Am J Gastroenterol.* 2006;101:S429–S430.
14. Cooney RM, Warren BF, Altman DG, et al. Outcome measurement in clinical trials for Ulcerative Colitis: towards standardisation. *Trials*. 2007;8:17.
15. Sandborn WJ, Regula J, Feagan BG, et al. Delayed-release oral mesalamine 4.8 g/day (800-mg tablet) is effective for patients with moderately active ulcerative colitis. *Gastroenterology*. 2009;137:1934–1943.e1–3.
16. 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;61:535–542.
17. Travis SPL, Schnell D, Krzeski P, et al. Reliability and initial validation of the ulcerative colitis endoscopic index of severity. *Gastroenterology*. 2013;145:987–995.
18. 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;62:201–208.
19. Rutgeerts P, Van Assche G, Sandborn WJ, et al. Adalimumab induces and maintains mucosal healing in patients with Crohn's disease: data from the EXTEND trial. *Gastroenterology*. 2012;142:1102–1111.e2.
20. 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;317:1625–1629.
21. Travis SPL, Higgins PDR, Orchard T, et al. Review article: defining remission in ulcerative colitis. *Aliment Pharmacol Ther.* 2011;34:113–124.
22. Feagan BG, Rutgeerts P, Sands BE, et al. Vedolizumab as induction and maintenance therapy for ulcerative colitis. *N Engl J Med.* 2013;369:699–710.
23. Vermeire S, O'Byrne S, Keir M, et al. Etrolizumab as induction therapy for ulcerative colitis: a randomised, controlled, phase 2 trial. *Lancet*. 2014;384:309–318.
24. Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *BMJ.* 1995;311:1145–1148.
25. 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.
26. FDA. Summary Minutes of the Oncologic Drugs Advisory Committee Meeting July 24, 2012 [Internet]. 2014. Available at: http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/Drugs/OncologicDrugsAdvisoryCommittee/UCM316323.pdf. Accessed January 7, 2015.
27. The Cancer Letter. FDA to Move Away from Central Radiology to Investigator Review in PFS Endpoint Trials—the Cancer Letter Publications [Internet]. 2012. Available at: http://www.cancerletter.com/articles/20120730. Accessed January 7, 2015.
28. Goldmacher G. *Decentralization of Imaging—Benefits and Caveats for Sponsors. Contract Pharma [Internet]*. 2012. Available at: http://www.contractpharma.com/issues/2012-11/view_features/decentralization-of-imaging/. Accessed January 7, 2015.
29. Dodd LE, Korn EL, Freidlin B, et al. Blinded independent central review of progression-free survival in phase III clinical trials: important design element or unnecessary expense? *J Clin Oncol.* 2008;26:3791–3796.
30. FDA. Guidance for Industry Standards for Clinical Trial Imaging Endpoints. Draft. [Internet]. FDA; 2011. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/UCM268555.pdf. Accessed January 7, 2015.
31. FDA. Guidance for Industry Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics [Internet]. 2007. Available at: http://www.fda.gov/downloads/Drugs/…/Guidances/ucm071590.pdf. Accessed January 7, 2015.
32. FDA. Guidance for Industry Developing Medical Imaging Drug and Biological Products Part 3: Design, Analysis, and Interpretation of Clinical Studies [Internet]. FDA; 2004. Available at: http://www.fda.gov/downloads/Drugs/GuidanceComplianceRegulatoryInformation/Guidances/ucm071604.pdf. Accessed January 7, 2015.
33. European Medicines Agency. Appendix 1 to the Guideline on Clinical Evaluation of Diagnostic Agents (CPMP/EWP/1119/98 REV. 1) on Imaging Agents [Internet]. European Medicines Agency; 2009. Report No.: EMEA/CHMP/EWP/321180/2008. Available at: http://www.ema.europa.eu/docs/en_GB/document_library/Scientific_guideline/2009/09/WC500003581.pdf. Accessed January 7, 2015.
34. 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 (Oxf).* 2014;2:201–206.
35. Rex DK. Looking over your shoulder during colonoscopy: potential roles for videorecording colonoscopy withdrawals. *Gastrointest Endosc.* 2012;75:134–137.
36. Ford R, Schwartz L, Dancey J, et al. Lessons learned from independent central review. *Eur J Cancer.* 2009;45:268–274.
37. Krzeski P, O'Leary DH, Zabbatino S. Lessons Learned from the Use of Central Endoscopy Review in Inflammatory Bowel Disease Trials [Internet]. Medpace; 2013. Available at: http://www.medpace.com/Offers/GI_CentralReader/Medpace_GI_Imaging_IBD_White_paper_May2013.pdf. Accessed January 7, 2015.
38. Daperno M, Comberlato M, Bossa F, et al; on behalf of IG-IBD Group. PC.01.8 increasing interobserver agreement on IBD endoscopic scoring systems: results from the IGIBDENDO educational program. *Dig Liver Dis.* 2014;46:S4.
39. 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.