Standardizing central reading for endoscopic assessment

Klaus Gottlieb, MD, MS, MBA
Senior Medical Director
Quintiles

Clinical Professor of Medicine
George Washington University
Washington, DC

GREAT3 March 30 to 31, 2015

Gastroenterology Regulatory Endpoints and the Advancement of Therapeutics

Silver Spring, MD
Dr. Gottlieb is an employee of Quintiles, a company that provides outsourcing services and consulting to the biopharmaceutical industry.

The views and opinions expressed in this presentation are his own and do not necessarily represent those of Quintiles and must NOT be taken to provide official policy or guidance on behalf of the FDA.
Standardizing central reading for endoscopic assessment

• Which FDA guidance documents are relevant?
• Is central reading of endoscopy (CROE) always beneficial?
• Should we use blinded central readers that decide independently or consensus panels?
• How many central readers should we use? Is one enough?
• How can we properly “average” ordinal data from more than one central reader?
• If we use more than one reader, should we involve the site reader? Why or why not?
• Can we mix and match read algorithm, for example for eligibility and efficacy?
• Do we need to do more work on inter-rater agreement?
• What do FDA, the academic community and industry need to do next?
## FDA Guidance Documents

*That address central reading in general*

<table>
<thead>
<tr>
<th>Guidance</th>
<th>Selected Key Message</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guidance for Industry: Standards for Clinical Trial Imaging Endpoints</td>
<td>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.</td>
</tr>
<tr>
<td>(Draft Guidance) (2011) updated March 2015, now titled “Clinical Trial Imaging Endpoint Process Standards”</td>
<td></td>
</tr>
<tr>
<td>Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics (2007)</td>
<td>Centralized independent verification of tumor endpoint assessments (especially for Progression Free Survival 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.</td>
</tr>
<tr>
<td>Guidance for Industry: Developing Medical Imaging Drug and Biological Products Part 3: Design, Analysis, and Interpretation of Clinical Studies (2004)</td>
<td><strong>Image interpretation is inherently subjective.</strong> Therefore, inter-reader 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.</td>
</tr>
</tbody>
</table>
Does Central Reading of Endoscopy (CROE) always work?  
*Two central reader studies – for better or for worse*

Central Read Gain- △ Effect Size  
+5.8%  
Oral mesalamine  
**Endpoint**: endoscopic component of Mayo Clinic Score (MCS)

Central Read Loss- △ Effect Size  
-20.1%  
Kobayashi et al. 2014 et al. 2014  
Rectal mesalamine  
**Endpoint**: endoscopic component of MCS
# The two trials side-by-side

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Compare central reader with site reader</td>
<td>Compare central reader with site reader</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Drug</th>
<th>mesalamine - oral</th>
<th>mesalamine - rectal</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Population</th>
<th>Mild-Mod</th>
<th>Mild-Mod</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Design</th>
<th>Phase 3 DB-MC-PCT</th>
<th>Phase 3 DB-MC-PCT</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Number of patients</th>
<th>281</th>
<th>129</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Endoscopic Endpoint</th>
<th>Modified endoscopic subscore of the MCS (friability binary)</th>
<th>Endoscopic subscore of the MCS</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Image capture</th>
<th>Video</th>
<th>Still photos</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Site Reader</th>
<th>decides enrollment, primary efficacy endpoint</th>
<th>decides enrollment, primary efficacy endpoint</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Central Read Model</th>
<th>single expert central reader</th>
<th>7 central readers with complex 2-stage adjudication process that progresses from ‘agreed group’ to ‘review meeting group’ (consensus process) as needed.</th>
</tr>
</thead>
</table>
What can we learn from Feagan and Kobayashi?

- No firm conclusions possible but some important hypotheses
  - Blinded central readers (not members of a consensus panel) using video recordings may be beneficial (Feagan et al.)
  - Consensus panels that decide the final read score jointly based on endoscopic still images may not be the best approach (Kobayashi et al.)

- Is the above against the Central Limit Theorem? More readers – improved accuracy?
  - No. A consensus panel is not an arithmetic average of a continuous variable. In fact, vocal members of the panel can bias the entire group

- Is “one central reader is enough”; adding more doesn’t help?
  - This can be approached empirically. Caveats are that inter-reader agreement does not equate with accuracy. Accuracy calculations require a gold or reference standard – not available
  - Every empirical study will be fully applicable only to the specific pool of central readers used: “Individual results may vary”.
  - Is there a way to look at this using statistical theory for more generalized insights?
How can we ‘average’ scores (ordinal data)

• Problem: The arithmetic mean of ordinal data is highly problematic

• Solution: Aggregate ordinal data in a way that allows further manipulation: count votes

• Juries average across binomial decisions
  › Voting increases the chances of being fair. Fair = accurate
  › Voting is counting binomial decision: yes or no.

• Application of the binomial concept is attractive because it can be extended all the way from ‘micro-decisions’ a rater makes in arriving at a score (for example, is this an ulcer, is this erythema, is this friability?) to the ultimate ‘knowledge aggregation’ of several central readers. The majority vote is the verdict

• The increase in accuracy of juries (over a single voter) can be modeled by the following binomial formula, aka the Condorcet Jury Theorem

• For a detailed discussion and application see BMC Medical Imaging 2015, 15:6

\[
P_c = \sum_{k=\frac{n+1}{2}}^{n} \binom{n}{k} p^k (1 - p)^{n-k}
\]
No reader can have zero-classification error

Power plummets with modest decreases of probability of being correct

Worked example for sample sizes n0 for control, n1 for active, mu0 for placebo response rate, mu1 for active response rate.

- Assume p correct classification for average reader is 0.8 rather than 1.0 (unattainable):
- With only one reader, test power decreases to 0.55 (blue path)
- Using 2+1 readers (k=3) and letting them vote brings power back up above 0.8 (red path)
## Should the investigator/endoscopist score?

<table>
<thead>
<tr>
<th>Let’s ignore the endoscopist</th>
<th>Let’s involve the endoscopist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endoscopists are a major source of bias, especially upon eligibility determination</td>
<td>Endoscopists who are aware that central readers will challenge upcoding will not engage in this practice (observer effect)</td>
</tr>
<tr>
<td>The average investigator/endoscopist does not want to learn endoscopic scoring systems</td>
<td>Investigators take pride in advancing science and understand why they are asked to receive training in the scoring system.</td>
</tr>
<tr>
<td>Source video quality is independent of whether endoscopist scores or not</td>
<td>Endoscopists who also score record videos from the perspective of what is needed by central reader which results in better quality.</td>
</tr>
<tr>
<td>Central readers are better readers. Acquiring requisite skills takes a long time.</td>
<td>Competent endoscopists adequately trained in the respective methodology perform at the level of central readers.</td>
</tr>
<tr>
<td>We save money by using just one or at most two readers.</td>
<td>We save money by improving study power.</td>
</tr>
</tbody>
</table>
Is it OK to have two different read algorithms in the same trial?

The change of the measurement instrument makes any change from baseline to first efficacy assessment hard to interpret.
In clinical trials do kappa decreases mean the raters are getting worse? Relevance to CROE quality control.

- Most research on rater agreement done in the context of one-time tests
- There is a previously not described dynamic behavior of kappa in clinical trials when inter-rater statistics are done repeatedly
  - Fluctuations occur that are independent of the underlying true rater performance
  - Caused by the methodology of the kappa calculation.
- Kappa is a function of the probability of disagreeing with a category, and this probability increases with narrower windows, but is also affected by the distribution of the expected counts
- In the case of a typical IBD clinical trial the distribution of the affected counts changes as the trial progresses: For example, in the vedolizumab phase 3 trial the average MCS was 2.4 at screening and then decreased at week 6 to 2.0 for placebo and 1.7 for vedolizumab
- Kappa will vary as the study progresses, either increasing or decreasing depending on the range of the data.
  - kappa will be lower when mean ranges of the MCS are closer to 1.5,
  - kappa will be higher for MCS close to either 0 or 3

What’s next?

• Continue empiric research and development of theoretical foundations including reader agreement statistics

• Large multicenter trials using different read algorithms are currently in progress

… but we can’t wait for the results because we need to write charters for new trials that are starting now

• Main questions center around the number of central readers and whether the site endoscopist should also be a reader
  › Both approaches have their proponents

• The concept of ‘adjudication’ means different things to different people, it is important to be very specific about what that means or avoid the term altogether, e.g., voting concept

• Industry, academia and regulators realize that CROE topics need to be discussed
  › Thanks for the opportunity to speak at GREAT 3!
  › Meeting at DDW 2015 about CROE being organized, regulators invited
  › Review paper about CROE submitted to Inflammatory Bowel Diseases
References


• Reeve R, Gottlieb K, Hussain F. “Inter-rater kappa can change during trial even when raters do not”. DIA-FDA Statistics Forum April 2015. Accepted Abstract.