

The imaging core lab perspective on glioblastoma imaging and response assessment in clinical trials

Gregory V. Goldmacher and David Raunig

ICON Medical Imaging, Warrington, Pennsylvania

Corresponding Author: Gregory Goldmacher, PhD, ICON Medical Imaging, 2800 Kelly Rd., Warrington, PA 18976 (gregory.goldmacher@iconplc.com).

Multicenter clinical trials that include medical images as a key component of response assessment often involve an imaging service provider (a core laboratory or contract research organization) to collect images and often to provide independent assessments of treatment response. The brief discussion and recommendations provided here are not intended as a rigorous academic analysis but reflect the practical experience accumulated at one such institution, which has conducted the image collection and review for numerous glioblastoma trials, in every phase of drug development, encompassing over 4000 patients scanned at over 900 sites.

Keywords: clinical trial, glioblastoma, imaging, independent review, RANO.

Image Acquisition

Standardization of images across sites is a significant challenge for reproducible assessments in a large multicenter trial. To get consistent data across diverse regions and capabilities, multicenter trials should have an imaging manual that specifies the acquisition parameters. Personnel at sites should be trained on the manual before the trial begins, and each scan should be reviewed soon after collection, with feedback to sites to enforce the specifications. Complex statistical models can compensate for some regional diversity but impose a cost (in degrees of freedom) on final study analyses, and it is better to avoid site-to-site differences if possible.

The scan acquisition requirements should be carefully planned before the study starts, because adding extra analyses later may not be possible. One example of a good idea introduced too late comes from the sponsors of one recent trial who decided, after approximately half the data were collected, to get information on enhancing tissue volume. Unfortunately, almost half of the scans performed did not include 3D acquisitions with thin slice reconstruction before and after contrast, rendering such assessment unreliable in these cases. Typical standard-of-care scan protocols do not include sequences that allow for analyses that might be desired in a trial, such as T1 subtraction or volumetric tumor assessment.

Consistent scan acquisition between visits is required for valid visit-to-visit comparisons. Even within a single institution, there are often multiple scanners with different manufacturers, field strengths, etc, and it is quite common to see patient scans conducted on different scanners at different times and with different settings, confounding visit-to-visit comparisons. This problem can

be mitigated through a well thought out communication plan between investigators and MRI technologists. The technologists should be made aware of what clinical research studies are active at the site, and the imaging requirements for each study. Investigators must inform the technologists (or a radiologist who can then inform the technologists) when a patient is also a trial subject and which scanner and protocol should be used. This is more difficult when a site is involved in multiple studies, so the study sponsor and the imaging contract research organization may need to become actively involved in ensuring site compliance. One of the most effective approaches to improve protocol compliance is to enlist a radiologist at the site as a coinvestigator, to oversee the process and be accountable for scan quality.

Transmission of image data to independent facilities has become faster and easier over the last few years, as more sites become capable of electronic data submission. Electronic submission has significant advantages over sending data on physical media when rapid central review is needed. There are both custom-built solutions and commercial providers of electronic submission services, and these should be used as often as possible, though this may require discussions with hospital information technology departments. Other imaging techniques, such as T1 subtraction, diffusion weighted imaging, perfusion, and spectroscopy, often require postprocessing. Postprocessing can be done in standardized ways, using commercially available software, and there is generally no difficulty handling data from multiple manufacturers. However, if postprocessing is done at an independent facility, sites must be trained to submit complete datasets, with all of the information required for postprocessing (such as b-values for diffusion or data to determine temporal

Received 2 June 2014; accepted 10 August 2014

© The Author(s) 2014. Published by Oxford University Press on behalf of the Society for Neuro-Oncology. All rights reserved.

For permissions, please e-mail: journals.permissions@oup.com.

resolution for perfusion) either embedded in the DICOM (Digital Imaging and Communications in Medicine) headers or attached separately.

Image Interpretation and Response Assessment in Neuro-Oncology Criteria

Presentation of imaging and clinical components as a single process combines the specialized expertise of radiologists and oncologists. Neuroradiologists have superior experience in reading the MRI scans, while neuro-oncologists are better able to interpret changes in steroid dose and clinical status. In our experience, the best approach is a 2-stage process, in which radiologists read the MRI scans, and then neuro-oncologists combine the results of the read with clinical information to give the final response. The Response Assessment in Neuro-Oncology (RANO) publication defined the categories of response but did not describe a step-by-step process for performing the analysis. In our experience, an algorithmic process for assessing scans, analogous to the process used for other oncology trials, provides standardization needed to reduce variability of interpretation. Such a process is even more valuable when interpretation is done by a large group of local readers with varying levels of experience in clinical trial image review.

The recommended interpretation process works as follows:

- At baseline, a radiology reviewer locates lesions that are amenable to quantification (measurable lesions), from which he or she selects the lesions that will actually be quantified at every visit (target lesions). All other disease is documented at baseline and followed qualitatively (nontarget lesions).
- At each visit after baseline, the radiology reviewer measures the target lesions, visually assesses the enhancing nontarget lesions and the T2/fluid attenuated inversion recovery (FLAIR) images, and searches for new lesions.
- The target lesion, nontarget lesion, and new lesion assessments are combined using a logic table (embedded within the case report form collection software) to produce a suggested overall response for the visit. Readers can agree with the calculated response or provide their own.
- After all visits are completed, the radiology reviewer can adjust earlier assessments (though not the baseline selection of target lesions) based on later data, documenting the reasons for any change in her or his response decisions. In particular, this allows an opportunity to assess whether early increases in enhancement represented true progression or pseudoprogression, and whether new findings that were initially questionable turned out to be new lesions when followed over time.
- After the radiological assessment of all the visits is complete, a neuro-oncologist combines the imaging results and the available clinical data in a similar manner, with individual visit assessments followed by a summary assessment during which earlier visit responses can be overridden. For this portion of the read, it is important to standardize how different trial sites collect clinical data (for example, do they record steroid dose changes from baseline or from the prior visit, or do they simply record the actual dose the patient is on?).

The process described above is relatively easier to carry out at an independent review facility once the patient has completed all

trial visits (typically after progression or death), and all visits can be reviewed in a single session by reviewers from a small pool of highly trained radiologists and neuro-oncologists. Pseudoprogression is a critical issue in this evaluation. Independent radiologists reading scans after the patient is off trial can avoid mistaking pseudoprogression for true early progression by not calling progression in a radiotherapy-treated area within 12 weeks of conclusion of radiotherapy, unless follow-up scans confirm that progression began within that 12-week period. It must be noted that in most trials, information clearly defining the radiotherapy-treated area is not collected and submitted to central read facilities, though ideally it should be.

The evaluation is more difficult when decisions have to be made in real time at trial sites, in particular because of the uncertainty around early increases in contrast enhancement. If a patient is deemed to have progressed prematurely and is taken off study during the pseudoprogression period, follow-up MRIs may be discontinued and the patient's data must be censored in the analysis. There is no perfect approach to this decision, but close communication among site radiologists, radiation oncologists, and neuro-oncologists is required. The availability of detailed information from clinicians improves the decision process surrounding pseudoprogression, but there is also the potential for investigators to unblind or bias radiology reviewers. Even in the setting of randomized, double blinded, placebo controlled trials, local reader bias is possible because medications have side-effect profiles that may lead to functional unblinding of investigators. This, along with the inconsistent training of local reviewers in the application of RANO, suggests a need to conduct comparisons between local and independent interpretations, to guard against bias and variability and to assess the impact of expert review on trial results.

Reader Performance Evaluation

If the radiological read uses what is referred to as a “2 + 1” design, common in late-phase trials, 2 primary radiologists assess each case, and defined disagreements are resolved, or adjudicated, by a third radiologist, who selects one of the primary readers' assessments as closest to correct in his or her judgment. This model of 2 primary radiology reviewers and an adjudicator has been used in several large late-phase glioblastoma trials. In addition to improving the read accuracy, it allows assessment of reader performance in the form of an interreader disagreement rate, also known as the adjudication rate.

Adjudication is typically triggered by disagreement on the date of progression (DOP), though other disagreements can also be adjudicated, especially in trials where response rate is the key endpoint. The rate of disagreement on DOP for GBM can be high relative to other cancers. As a general rule, independent readers disagree on the DOP 40%–50% of the time, though roughly half of these disagreements fall within one visit of each other (this is based on the authors' experience with several large trials; preparation of data for publication is in progress). The high disagreement rate reflects the difficulty of agreeing on the exact visit on which progression occurred. It does not indicate that reader assessment of progression is the equivalent of a “coin flip,” which would ignore reader competence and imply that any study visit would be equally likely to progress.

The most significant single source of disagreement is the evaluation of disease seen on T2/FLAIR images. In our experience,

almost half of the disagreements resulted when one reader called progression on the FLAIR images at a particular visit, while the other reader did not. This disagreement was especially prominent in trials where an anti-angiogenic agent was part of the treatment. Another major source of variability is readers' differing willingness to call a new lesion. Finally, variability arises from decisions regarding where to measure lesions in the first place, especially when lesions are highly irregular and contain necrosis, cystic change, and so forth.

Evaluating the performance of the final read requires a much closer review of the details of the assessments than just the disagreement rate. Prior regulatory opinions are cited as indicating that a disagreement rate above 40% indicates unreliable data, but a great deal of glioblastoma trial experience shows that such rates are common, while a thorough review of the data shows no evidence of poor data quality, given the difficulty of the disease assessment. While the absolute disagreement rate should be reported, the rate of disagreement within one visit should always accompany those statistics. A thorough and extensive analysis of the disagreement on the date of progression strongly points to evaluating for evidence of single reader bias as a much more appropriate analysis of reader performance than adhering to a strict disagreement rate threshold.

Another important source of data for determining the source and import of various imaging features is the adjudicator comments. Adjudicators must be trained to provide comments that are as clear and specific as possible regarding the disagreement between the readers and their reason for choosing their preferred interpretation. Searchable terms are recommended to facilitate

the use of software to compile the comments, though there is no substitute for a thorough review of the comments to get a comprehensive understanding of reader performance.

Key Recommendations and Observations

- Prespecify the evaluations to be carried out, design the scanning protocol to support these analyses, and distribute the image acquisition requirements to sites in an imaging manual to all of the sites to ensure uniform image quality.
- Conduct ongoing near real-time review of collected images to provide timely feedback to sites about quality.
- Involve a local radiologist as a coinvestigator, if possible, and establish clear communication channels with the technologists.
- Response assessment should include both radiologists and neuro-oncologists and should be done in a defined algorithmic manner to reduce interpretation variability. Both radiologists and neuro-oncologists should be experienced in the application of the RANO criteria.
- The greatest single source of disagreement between readers is progression of disease based on T2/FLAIR images.
- Using experienced, multiple readers for each case can increase assessment accuracy. However, using the rate of disagreement by itself is inadequate to evaluate reader performance given the difficulty of interpretation in this indication. Statistical methods more sensitive and specific to reader performance quality assessment should be used to evaluate reader performance more effectively.