CALYX™

Reviewer Performance Monitoring in Blinded Independent Central Review Setting

Manish Sharma, Michael O'Connor, Nicholas Enus, A. Kassel Fotinos-Hoyer, Sayali Karve, Yibin Shao, and Oliver Bohnsack

CALYX RATIONALE FOR INDEPENDENT REVIEW AND COMMON PARADIGMS

Utilizing an established imaging evaluation framework in oncological clinical trials enables use of surrogate imaging-based primary endpoints such as progression-free survival (PFS), time-to-progression (TTP), and/or objective response rate (ORR) in lieu of an overall survival (OS) endpoint. As opposed to OS however, surrogate imaging-based endpoints (e.g., PFS, TTP, and ORR) in oncological clinical trials may be subject to variations due to differences in image acquisition and assessment procedures. Therefore, evaluation of clinical indicators and assessments involving radiological images is a challenging task where the criteria for determining quality and/or precision of the imaging assessment is of utmost importance. Assessment of imaging data in support of surrogate endpoints by an independent entity reduces evaluation bias and improves assessment consistency.

A blinded independent central review (BICR) approach, where independent radiology reviewers are managed by a central lab, assess subject images for a specific trial, and are blinded to clinical patient data and treatment decisions, is recommended by the United States Food and Drug Administration (FDA) for registration of oncology trials.[1,2] During BICR, independent radiology reviewers provide an assessment of subject imaging independent of on-site clinical trial research investigator(s) and are blinded to patient information (e.g., name, date of birth), assessments made by the investigator, randomization arm, the total number of imaging timepoints available for each subject, and various other potentially biasing non-radiological information.

Depending on the anticipated subjectivity of the review type and depending on the planned endpoint analyses, various review models are used in the BICR approach (e.g., single-read, double-read with adjudication, etc.).

Currently, Response Evaluation Criteria in Solid Tumors version 1.1 (RECIST 1.1) guidelines are used for assessing changes in overall tumor burden in oncological clinical trials through initial selection of a small number of representative "target" lesions that are chronologically followed and measured over time. RECIST evaluation focuses on important aspects of assessment such as standardization of the tumor response assessment process by categorization of the tumor burden, identification and selection of defined numbers of localized target lesions, minimum tumor size, measurability, and acceptable frequency and separation of assessments.

VARIABILITY IN INDEPENDENT REVIEW ASSESSMENTS

Assessing the same subject images by multiple reviewers will lead to some discrepancy in the overall assessment of a subject based on the radiographic images.[3] Many factors influence discrepancy rates among independent reviewers in the assessment of subjects' radiographic images using assessment criteria such as RECIST 1.1.[4] Acknowledging this, registration oncology trials commonly use the highly advocated double read with adjudication review model when investigating efficacy and reviewer performance. In this review model, two well-qualified, board-certified radiology reviewers, also referred as "readers" (double read with adjudication model), each assess subject images independently of one another. Then, in the case of a discrepancy in their assessments, a third radiologist, referred to as the "adjudicator," selects one of the two radiologists' assessments in their entirety, according to who the adjudicator agrees with most. Therefore, overall monitoring of adjudication rates is highly advocated throughout the trial to monitor reviewer performance.[5,6,7,8]

INDEPENDENT REVIEWER PERFORMANCE MONITORING METHODS

A number of indicators are used to monitor independent reviewer performance, but not all are created equally, and it is therefore necessary to monitor a combination of indicators in order to achieve a well-rounded understanding of reviewer performance.

1. ADJUDICATOR RATE

Adjudication rate (AR) is the most frequently used indicator when evaluating independent reviewer performance and trial efficacy. AR is defined as the number of cases triggered for adjudication divided by the total number of all cases read, in %, as shown in Equation (1):

$$\text{Adjudication rate} = \frac{\text{\# of cases triggered for adjudication}}{\text{total \# of cases read}} \times 100 \quad (1)$$

It is implied that AR is inversely related to the quality of BICR data and to the quality of image assessment provided by the independent reviewer. However, AR does not consider how often the adjudicator agreed with a given reviewer (i.e., the adjudicator agreement rate) or the inverse (i.e., the adjudicator disagreement rate). AR also does not consider the cases where no adjudication is required. A high AR may not always indicate poor reviewer performance for a given reviewer if also associated with a high adjudicator agreement rate, and a low AR may not always indicate good reviewer performance for a given reviewer if associated with a high adjudicator disagreement rate.

Understanding the underlying cause(s) of inter-reviewer and intra-reviewer variability is important to minimize the AR. Improper understanding of the cause(s) may lead to wasted resources and may increase Type I errors (false positives), raising regulatory concern. An AR >30% is typically considered to be an indicator of poor reviewer quality. However, AR is influenced by a number of factors such as interpretation or procedural errors, differences in the subjective assessment of total tumor burden, and differences in determination of the most representative lesions which are also suitable for repeat measurements. **The major drawback of AR is that it does not consider how often the adjudicator agreed with a given reader, which is a critical consideration for assessment of reader quality for both primary readers.** Therefore,

AR as a metric of reviewer quality should be carefully used and interpreted and should not be the sole indicator of reviewer performance.

2. ADJUDICATION AGREEMENT RATE

The adjudicator agreement rate (AAR) is a relative performance indicator for a given reviewer as compared to the other reviewers a given study, with a higher adjudicator agreement rate suggesting better reader performance.

$$\text{Adjudicator agreement rate (AAR)} = \frac{\text{\# of cases where adjudicator agreed with given reader}}{\text{total \# of all cases adjudicated}} \times 100 \quad (2)$$

AAR as shown in Equation (2) above is a more reliable measure of individual reviewer performance as compared to the AR alone. Since AR does not consider the total number of cases adjudicated for a given reader, it may incorrectly identify poor performers. AAR is less prone to the same issues since it is directly related to reviewer performance.

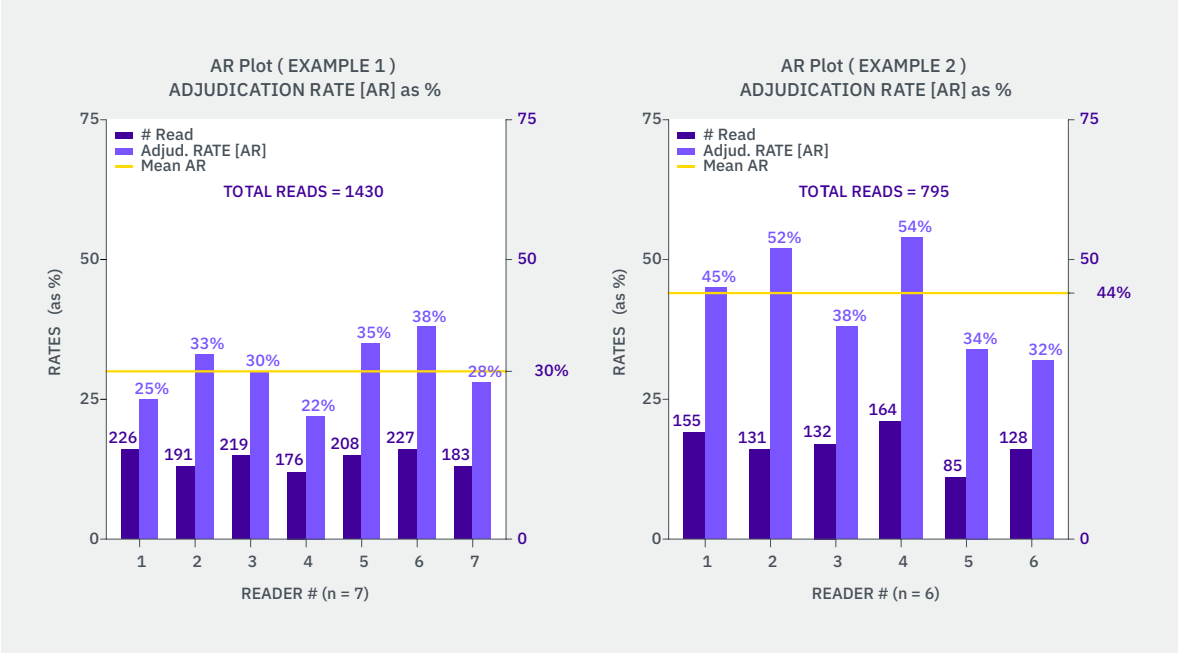

Figure 1: Sample AR plots showing adjudication rate for the study & all readers

However, low AAR may not always indicate poor reader performance if associated with a low AR. Furthermore, AAR does not include the total number of cases adjudicated for a given reader. Thus, it may also incorrectly identify poor performers.

3. READER DISAGREEMENT INDEX

Reader Disagreement Index (RDI), an improved and innovative indicator to monitor individual reader performance, takes into account both the overall AR for a study as well as the individual AAR for each reviewer. The RDI indicates the percentage of disagreed cases for a given reader across the total number of cases read, as defined in Equation (3) where a low RDI value indicates better reader performance and a high RDI value indicates poorer reader performance.

$$\text{RDI} = \frac{\text{\# of cases where adjudicator disagreed with given reader}}{\text{total \# of all cases read}} \times 100 \quad (3)$$

RDI combines aspects of both AR and AAR in that it considers the subjects for which adjudicator disagreed with the reader (as shown in numerator of Equation 3) and also considers adjudicator disagreement relative to the total number of cases read (as shown in the denominator for Equations 1 and 3).

RDI is represented by its mean ± standard deviation (SD) value. Mean or average is a summarizing statistic and measure of the center of a distribution. Mean is calculated as a sum of the observations divided by the number of observations. It is particularly meaningful if the data are symmetrically

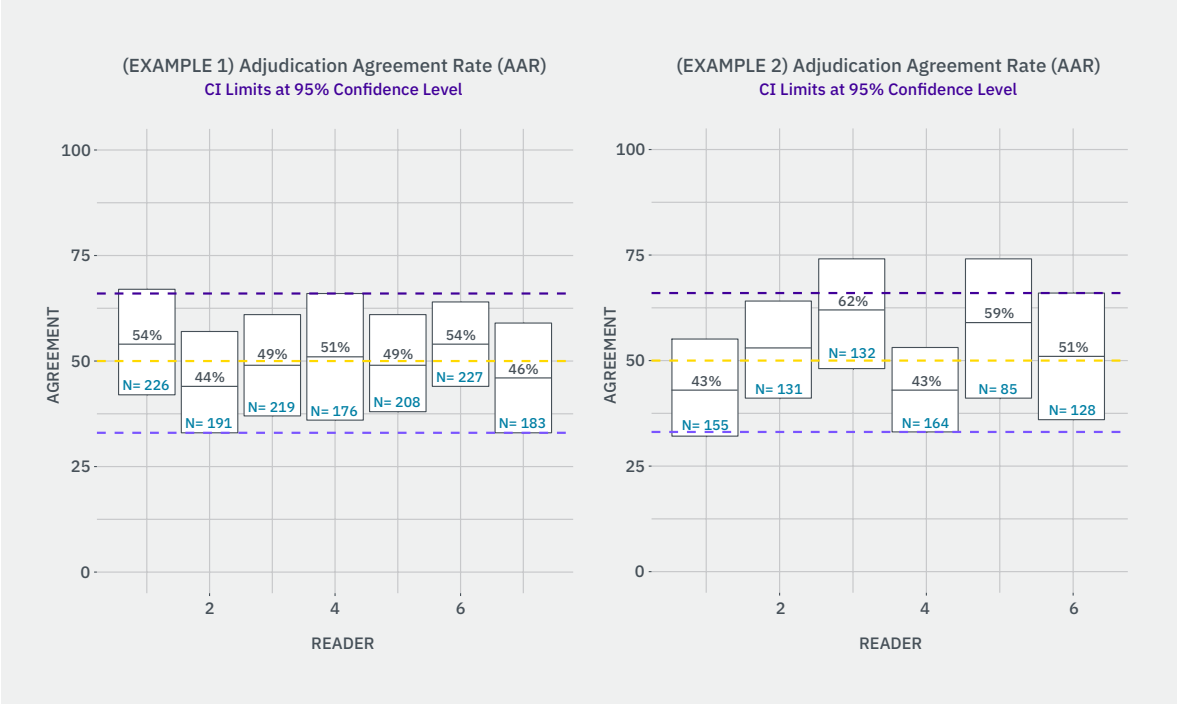

Figure 2: Sample AAR plots showing adjudication agreement rate for all readers

distributed below and above the average or mean. When considering a dataset, use of SD, which measures the spread of the data, is a must. The larger the magnitude of the SD, the greater spread of the data exists.[9] It is recommended to indicate mean RDI with 2-SD.

CASE STUDY 1:

M. Sharma et al.[10] performed a retrospective review of adjudication data for 20 oncology clinical trials with a total of 7163 subjects (32,536 timepoints) assessed using RECIST 1.0 or 1.1 at ASCO 2018. AR, AAR, and RDI were generated per reader per study. RDI identified the discordant reader in all 20 studies, whereas AR and AAR identified the discordant reader in 13 and 12 of the 20 studies, respectively. In 3 studies, the reviewer with the highest % adjudicator disagreement had neither the highest AR nor lowest AAR. This reviewer could have been missed without the RDI indicator. RDI proved to be an effective quality indicator by combining AR and AAR to identify potential outliers and an excellent tool for identifying the discordant reviewer.

Figure 3: Sample RDI plots showing adjudication agreement rate for readers with mean+/-2SD

CASE STUDY 2:

M. Sharma et al.[11] presented a retrospective analysis of 3 oncology clinical trials used to study the discordance between two reviewers using RECIST 1.1 assessments in a BICR setting. Broadly, 10 board-certified radiologists reviewed three studies at a central site, with an average of 5 reviewers per study. The study included data of 349 subjects and 1361 total timepoints (beyond baseline). RDI was calculated to identify the reviewer with the highest level of cases disagreed by the adjudicator (i.e., the discordant reviewer) along with calculation of AR and AAR.

CASE STUDY 3:

M. Sharma et al.[12] performed a detailed review of BICR adjudication data for 12 oncology clinical trials, with a total of 5369 subjects (ranging from 119 to 894 per individual study) with 27,056 timepoints assessed using RECIST 1.0 or 1.1, the Lugano classification or iwCLL assessment criteria. The results were shared at ASCO 2019. RDI for each reviewer was used to identify the discordant reviewer (i.e., reviewer with the highest level of cases disagreed with by the adjudicator) when approximately 10% of the total reads were completed for each study. RDI was also compared with AR and AAR on an ongoing basis throughout the study. Mean RDI + standard deviation (SD) were used to identify outlier reviewers. RDI reliably identified the most discordant reader consistently across all 12 studies, while AR & AAR did not. The results confirmed the advantage of RDI as a leading indicator for independent reviewer performance across indications and criteria using a double read with adjudication review model. RDI, when calculated as early as at the 10% of total reviewed cases benchmark, demonstrated a positive predictive value of 91% and negative predictive value of 93% (Sensitivity 71%; Specificity 98%). Early identification of an outlier reviewer as per RDI (i.e., after reviews completed for ~10% study

visits), followed by detailed analysis and corrective measures, such as retraining the reviewer, can serve as timely intervention to improve review quality.

	OUTLIER – YES (100%)	OUTLIER – NO (100%)	
Outlier – Yes (10%)	10 (TP)	1 (FP)	11 (at 10%)
Outlier – No (10%)	4 (FN)	53 (TN)	57 (at 10%)
	14 (Outlier)	54 (Non-Outlier)	68 Total

4. INTRA-REVIEWER AND INTER-REVIEWER VARIABILITY

Research has reported that given the same task on the same imaging data, there can be considerable inter-observer variance between two radiologists.[13] A summary of radiological observer/reviewer variability research has been outlined by Manning.[14] Monitoring the reproducibility of individual reviewer performance (intra-reviewer variability) and the consistency of reviewer performance across all study reviewers (inter-reviewer variability) during a trial is highly recommended. Upon evaluation, corrective actions such as retraining reviewers may be required. However, without understanding of underlying cause, retraining activities may be a waste of time and resources

With respect to clinical trials, even with clear BICR procedures, the FDA recognizes that there will be assessment variability[15] and therefore recommends that "trial developers should consider the potential effect of reader interpretation variability upon the clinical trial outcomes." The same FDA Guidance document instructs that "The charter should describe the process for monitoring compliance with the image display and interpretation process." One fundamental factor in monitoring variability is the so-called "adjudication trigger." A study design might specify adjudication for any discordant assessments between two reviewers or a more selective adjudication "trigger" may be specified (e.g., Progression of Disease (PD) assessment versus any other assessment). For this reason, variability between studies can be due to study design as well as by the disease indication under review in the study as has been reported by Ford et al.[16]

DISCORDANT ASSESSMENT MONITORING

Development and usage of additional monitoring methods can particularly provide further insight into an individual reviewer's performance. By performing more specific analyses, we can "flag" discordant assessment pairs in which an individual reviewer could have a statistically significant lower AAR per type of assessment.

1. REASONS FOR DISCORDANT ASSESSMENTS

Several reasons for discordant assessments include aforementioned inherent variability between radiologists and understanding of the study protocol. Regarding variability between radiologists, lower-than-expected AAR (e.g., a partial or complete response assessment type) could be "bias signal" with respect to evaluation of patients' response to treatment. Misunderstanding of the protocol might explain a reader's lower AAR which could be particularly useful early in a clinical trial. An example might include a lower AAR, between two response assessments (i.e., partial versus complete). While such a difference would likely have no effect on endpoint determinations, it might indicate that some protocol specification (e.g., a measurement threshold) is not clear to a radiologist.

2. ANALYZING A DISCORDANT REVIEWER

In case of reviewer response or progression biases, a reviewer "flag" will play an important role and

can be used to determine a need for training and/or independent review charter (IRC) clarification. A retrospective analysis of dual-assessment variability was studied by J O'Connor et al., in which dual assessments were made for 704 subjects by 6 board certified radiologists.[16] One reviewer had a low overall AAR compared to the other five reviewers. A specific evaluation for that reviewer was required that warranted an analysis of all discordant assessment pairs. An automated evaluation function would facilitate analysis where all reviewers in the study were scanned to monitor whether any reviewer's pairwise discordant assessments revealed a probability of a lower-than-expected AAR for discordant pairs. This might not be visible by examining only the overall AAR value. The fundamental approach was to analyze each reviewer's assessment pairs. In the example study, there were five RECIST 1.0 or 1.1 assessment codes defined in the IRC. The abbreviations and meaning of codes for this study are listed in Table 1.

	ABBREVIATIONS	MEANING
1	NE	Not Evaluable
2	CR	Complete Response
3	PR	Partial Response
4	SD	Stable Disease
5	PD	Progressive Disease

Table 1: Summary of RECIST Assessment Codes in the Example Study

3. DISCREPANCY GRIDS OR ASSESSMENT-PAIR GRIDS

While assessment "certain-truth" is not known, we assume "quasi-truth" is determined by either reviewer agreement or, when there is assessment discordance, by the adjudicator's selection of the reviewer with whom he/she agrees with most. It is assumed that because all reviewers have similar board certification as well as training on and competence in understanding the trial's IRC that on average, adjudicators would likely agree with each reviewer in approximately 50% of the adjudication cases. What we refer to as a "bias" signal is a statistically significant probability of the AAR being less than 50% for each aggregated assessment pair. Each possible discordant assessment pairing that triggered adjudication was listed in a grid and the number of those pairings that the adjudicator agreed with per reader was assessed. This can be illustrated by presenting discrepancy grids or assessment-pair grids.

In Figure 4, a discrepancy grid or assessment-pair grid tabulates one of the six reviewers' assessments. The reviewer's (in this example Reader 1) assessments are tabulated (by row) and paired with the corresponding assessment by the other reviewer (by column). For each assessment pair, the other reader can be any one of the other five (not all of the other five).

Visual examination of the three purple-highlighted cells in the two matrices suggests that these pairs would have a much lower AAR than expected. This presents the question: do the apparent low AARs in these pairs have statistical significance? A binomial statistical test is used to give an objective answer. The binomial test implemented in R programming script augmented by RStudio performs an exact test of a null hypothesis (i.e., with input parameters set so that the probability of adjudication agreement is 50% and the alternative hypothesis that the true probability of agreement is less than 50%).[17] The confidence for the upper confidence interval (CI) level of the estimated agreement rate is also calculated. CIs for this R function are implemented in the manner of Clopper and Pearson.[18] Both the statistical test threshold (p-value) and the confidence level can be adjusted, of course, to change the result's sensitivity. The values used to determine the results for this illustration were p-value threshold = 0.05 and confidence level = 99%.

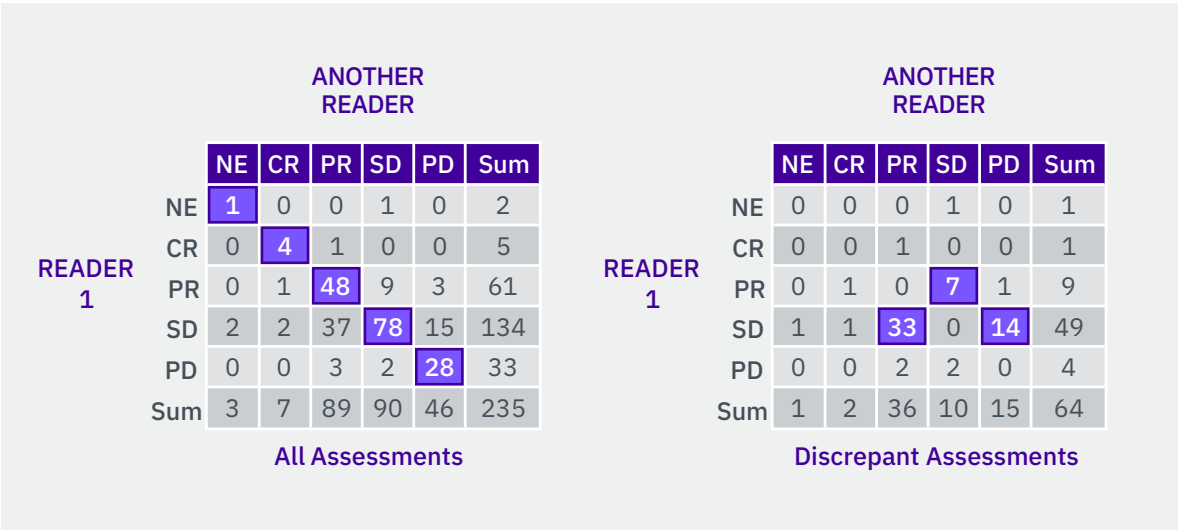

Figure 4: Reviewer 1's assessments (by row) versus another reviewer (by column). Purple-highlighted cells along the diagonal in the left matrix are matched time point assessments. LEFT: All assessment pairs; off-diagonal cells tabulate discordant assessments (i.e., adjudication pairs). RIGHT: All discordant assessments in which the adjudicator agrees with the other reader. The three purple-highlighted cells are of interest because of the seemingly large number of discrepant assessments.

The first two rows (adjudication pairs: SP/PR and SD/PD) result in rejection of the null hypothesis.

The last row is for an adjudication pair (PR/SD) in which the null hypothesis is accepted (no statistically significant probability that adjudication agreement is less than 50%).

Often, monitoring might be focused on low adjudication agreement that involves PR or CR or PD as these could impact endpoint determination. For illustration in Figure 4 and Table 2, there were two matrix cells among the three purple-highlighted cells for Reviewer 1 in which there were statistically significant probabilities (p-value < 0.05) that the reviewer's AARs for these specific cells were less than 50%:

- SD/PR (99% confidence of AAR for this cell being less than 28%)
- SD/PD (99% confidence of AAR for this cell being less than 37%)

In addition to monitoring discordant assessment pairs that include specific assessment codes such as just discussed, it can be useful to automatically analyse all matrix cells tabulating discordant pairs (an example would be a clinical trial where it may be useful to see if any cells have low AAR). As previously said, cells flagged for low AAR might indicate a reviewer's bias.

PAIR	TOTAL	DISAGREE	AGREE	ESTIMATE OF AGREEMENT	P-VAL	NULL-HYP	UPPER CONF AT 99% CONFID
SD/PR	37	33	4	11%	5.4e-07	REJECT	28%
SD/PD	15	14	14	7%	0.0005	REJECT	37%
PR/SD	9	7	2	22%	0.09	ACCEPT	66%

Table 2: Results of Binomial Test for the Four Highlighted Pairs in Reviewer 1 (Figure 4)

CONCLUDING REMARKS

RDI proves to be a more reliable quality indicator as compared to AR and AAR, as RDI can additionally identify the discordant reader, therefore improving its reliability. RDI offers advantages of identifying the most discordant reviewer that may be missed by analysis of AR and AAR alone for reviewer performance monitoring. Adding automated analysis of all or selected discordant assessment pairs for each reviewer in a study further improves the ability to monitor reader interpretation performance at a detailed level. Once a probability of low AAR has been "flagged," it would be prudent to further evaluate the signal. Discrepancy grids/assessment pair grids improve the capability to monitor BICR reviewers' performance in specific trials. These methods can be used to explore a reviewer having a low overall AAR or high RDI.

REFERENCES

1. Guidance for Industry Developing Medical Imaging Drug and Biologic Products. Part 3: Design, Analysis, and Interpretation of Clinical Studies. US Department of Health and Human Services. Food and Drug Administration. Center for Drug Evaluation and Research. Center for Biologics Evaluation and Research; 2004.
2. Clinical Trial Imaging Endpoints Process Standards Guidance for Industry Draft. US Department of Health and Human Services. Food and Drug Administration. Center for Drug Evaluation and Research. Center for Biologics Evaluation and Research. March 2015 Revision1.
3. Clinical Trial Imaging Endpoints Process Standards Guidance for Industry. US Department of Health and Human Services. Food and Drug Administration. Center for Drug Evaluation and Research. Center for Biologics Evaluation and Research; 2018.
4. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, et al. New response evaluation in solid tumors: Revised RECIST guideline (version 1.1). Eur J Cancer 2009; 45: 228-247.
5. United States Food and Drug Administration Guidance for Industry: Clinical trial endpoints for the approval of cancer drugs and biologics. Rockville, MD: US Department of Health and Human Services; 2007.
6. Dodd LE, Korn EL, Freidlin B, Jaffe CC, Rubinstein LV, et al. Blinded Independent Central Review of Progression Free Survival in Phase III Clinical Trials: Important Design Element or Unnecessary Expense? Journal of Clinical Oncology, 2008, 26(22), 3791-3796.
7. Cohen, K.L., Gönen, M., Ford, R.R., "Monitoring Reader Metrics in Blinded Independent Central Review of Oncology Studies," J. Clin Trials 2915, 5:4 (2015).
8. Ford RR; O'Neal M; Moskowitz SC; Fraunberger J (2016). Adjudication Rates between Readers in Blinded Independent Central Review of Oncology Studies. J Clin Trials, 2016, 6-5.
9. Pharmaceutical Statistics: Practical and Clinical Applications, Fifth Edition Sanford Bolton, Charles Bon, Informa health care, USA, 2010.
10. Manish Sharma, Anitha Singareddy, Michael O'Connor, A. Kassel Fotinos-Hoyer, Sayali Karve, Nicholas Enus, Daniel Clark, "Reader disagreement index (RDI) as an indicator of reader performance," Journal of Clinical Oncology 36, no. 15_suppl http://ascopubs.org/doi/abs/10.1200/ JCO.2018.36.15_suppl. e18592, Published online June 01, 2018; DOI: 10.1200/ JCO.2018.36.15_suppl.e18592.
11. Manish Sharma, J. Michael O'Connor, Anitha Singareddy, "Reader Disagreement Index: A better measure of overall review quality monitoring in an oncology trial compared to adjudication rate," Proc. SPIE 10952, Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment, 109520Q (4 March 2019); doi: 10.1117/12.2512611.
12. Manish Sharma, Oliver Bohnsack, Michael O'Connor, Yibin Shao, Nicholas Enus, Sayali Karve, A. Kassel Fotinos-Hoyer, "RDI as a method for reviewer performance monitoring in BICR setup for improving data quality," ASCO Journal of Clinical Oncology June 2019.
13. Kundel, H.L., Polansky, M., "Measurement of Observer Agreement," Radiology 228, 303-308 (2003).
14. Manning, D., "Cognitive factors in reading medical images: A survey of cognitive factors and models of medical image interpretation," The Handbook of Medical Image Perception and Techniques, (ed. E. Samei and E. Krupinski), Cambridge University Press, 2010.
15. Clinical Trial Imaging Endpoint Process Standards, Guidance for Industry (U.S. Dept. of Health and Human Services Food and Drug Administration, CDER and CBER), Final, April 2018.
16. J. Michael O'Connor, Manish Sharma, Anitha Singareddy, "Development of methods to evaluate probability of reviewer's assessment bias in blinded independent central review (BICR) imaging studies," Proc. SPIE 10952, Medical Imaging 2019: Image Perception, Observer Performance, and Technology Assessment, 109520P (4 March 2019); doi: 10.1117/12.2512603.
17. R function based on Conover, W.J., Practical nonparametric statistics. New York: John Wiley & Sons (1971). Pages 97-104.
18. Clopper, C.J. & Pearson, E. S. (1934). "The use of confidence or fiducial limits illustrated in the case of the binomial." Biometrika, 26, 404–413.

CALYX™

Reliably solving the complex.

calyx.ai
contact us at: hello@calyx.ai