Using reader disagreement index as a tool for monitoring impact on read quality due to reader fatigue in central reviewers

Sharma, Manish, Madasu, Madhuri, Kota, Sree Sudha, Bajpai, Surabhi, Shao, Yibin, et al.

Manish Sharma, Madhuri Madasu, Sree Sudha Kota, Surabhi Bajpai, Yibin Shao, Srinivas Pasupuleti, Michael O'Connor, "Using reader disagreement index as a tool for monitoring impact on read quality due to reader fatigue in central reviewers," Proc. SPIE 12035, Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment, 120350J (4 April 2022); doi: 10.1117/12.2613082

SPIE.

Event: SPIE Medical Imaging, 2022, San Diego, California, United States

Using "Reader Disagreement Index" as a tool for monitoring impact on read quality due to reader fatigue in central reviewers

Manish Sharma[1], Madhuri Madasu[1], Sree Sudha Kota[1], Surabhi Bajpai[2], Yibin Shao[3], Srinivas Pasupuleti[1], Michael O'Connor[4]
[1]Calyx.ai, Hyderabad, India, [2]Calyx.ai, Billerica, Boston, MA, US, [3]Calyx.ai, Shanghai, China, [4]University of Massachusetts, Lowell, MA, US

ABSTRACT

Purpose: Fatigue may lead to high medical errors by the radiologists. Fatigue is often described as feelings of weakness, lack of energy, and a desire to rest, and is associated with impairments in the ability to function. Visual fatigue has importance in medical imaging as errors (false-negatives) are relatively common. Blinded Independent Central Review (BICR) is a well-used method employed in many oncology registration trials. Ongoing monitoring of radiologist "reviewer" performance is both good clinical trial practice and a requirement by regulatory authorities. We use Reader Disagreement Index (RDI) as a potential tool to identify reader fatigue and compare reader fatigue in reviewers performing single versus multiple type of study.

Methods: A retrospective analysis of reviewers' RDI in four different clinical trials were performed. Fourteen reviewers' performance was analyzed with data for 3750 subjects having a total of 15105 timepoints across all clinical trials. These individual trial reviews were conducted by 14 board-certified radiologist reviewers using several established imaging assessment trial criteria. The objective of the study was to establish RDI as an effective tool to analyze if it could be a good surrogate marker for quality impacted by reader fatigue.

Results: The results indicate the RDI can be used as a tool to track reader quality which in turn may be able to predict reader fatigue. In the random pool of readers and studies analyzed, we did not notice any major trend or impact on read quality given that these trials were anyway actively monitored for read volume distribution and quality.

Conclusions: Fatigue may lead to high medical errors by the radiologists. RDI can be used as a good surrogate for read quality to monitor reader fatigue. Based on the results, it can be said that it is better to undertake more cases in a single study than undertake less number of cases in different types of studies to prevent reader fatigue.

1. INTRODUCTION

Medical errors by health care professionals are predicted to be 9.5% per year and forms the third most common cause of death in the United States. Among the radiologists, this rate can be as high as 30% of retrospectively detected errors (false-positives and false-negatives). Radiologists while reading and analyzing images as part of their clinical interpretation, should keep errors to as minimum as possible. Several factors can contribute to the error made by the radiologist which includes prolonged concentration, comprehensive image review, varied image loads, less time and more demands for quick read and complex anatomy represented by 3D visualization. Together, these factors contribute to fatigue. Fatigue is defined as a weariness and depletion of energy that can manifest physically and cognitively [1]. Fatigue may lead to high medical errors by the radiologists. Out of more than 11,000 preventable hospital deaths in English NHS acute hospitals each year, diagnostic error contributes to over eight thousand. It is known that more radiological errors are found towards the end of a shift then at the beginning, and longer duration of shifts in medicine are associated with higher rates. The situation is worsened due to increase in radiology workload without corresponding increase in the staff.

Fatigue is often described as feelings of weakness, lack of energy, and a desire to rest, and is associated with impairments in the ability to function [2]. Sometimes, the terms, sleepiness and fatigue are replaced with each other. Sleepiness is defined as drowsiness, sleep propensity, and decreased alertness. Fatigue is typically described as weariness, weakness, and depleted energy. Although the two conditions often coexist, fatigue can occur without sleepiness. Exertion and

Medical Imaging 2022: Image Perception, Observer Performance, and Technology Assessment, edited by Claudia R. Mello-Thoms, Sian Taylor-Phillips, Proc. of SPIE Vol. 12035, 120350J
· © 2022 SPIE · 1605-7422 · doi: 10.1117/12.2613082

discomfort are physical manifestations of fatigue and lack of motivation and "sleepiness" are considered mental components. Lack of energy reflects both physical and mental aspects of fatigue [3].

Fatigue is further divided into two subtypes: physical and mental fatigue. Physical fatigue involves deterioration in the muscles' ability to create or sustain force, which results in difficulties in controlling and coordinating muscles, while mental fatigue is a reduction in the ability to perform mental tasks. Fatigue can be measured by objective or subjective method, and each measurement method may cover one or more of these categories. The majority of these rely of subjective assessment in the form of self-reported rating scales or checklists. Krupinski and Reiner reported on several additional measures, including the Pittsburgh Sleep Quality Index and the Swedish Occupational Fatigue Inventory (SOFI). Objective measures of fatigue include the Maintenance of Wakefulness Test which measures alertness during the day. Here, the participants sit in a dimly lit room and are instructed to remain awake; the Psychomotor Vigilance Task for testing sustained attention, during which participants are instructed to respond to stimuli by pressing a button. The number of responses/failures to respond) are measured; and the Continuous Performance Test also measures sustained attention during which participants are presented with unstimulating tasks and are instructed to respond to certain stimuli while ignoring others. Visual fatigue has importance in medical imaging. Interpretation of medical images is a challenging task which is also repetitive in nature, and errors (false-negatives) are relatively common.

Besides, self-reporting, the other approaches to examine visual fatigue are- the assessment of accommodation (the adjustment of the curvature of the lens of the eye to change focus between objects in the distance and up close) and vergence (obtaining/maintaining binocular vision through the movement of eyes in opposite directions). Usually, accommodation and vergence decline with fatigue. Understanding how fatigue can result in missing clinically relevant abnormalities is important [2]. In addition to physiologic, emotional, and cognitive fatigue encountered throughout medicine, radiologists are also subject to visual fatigue due to prolonged periods of time interpreting complex medical imaging datasets on Computers [4].

Blinded Independent Central Review (BICR) is one of recommended method by the FDA for registration oncology trials [5-10]. An important end-point factor in oncology trials is the objective evaluation of clinical indicators involving radiological images. "Double read with adjudication" is one of the frequently used model in various oncology trials review. This method enables better radiological assessments than a single read model. Even though reviewers are always well qualified (board-certified and experienced radiologists), it is mandatory on trial management to monitor reviewer performance throughout the trial. Monitoring ensures high quality radiological assessments as well as adherence to specific clinical trial protocols. We have come to view Reviewer Disagreement Index (RDI) as a fundamental measure of reviewer performance.

Reader Disagreement Index (RDI) is an improved indicator for monitoring individual reader performance. It is an innovative indicator which takes into account both the overall adjudication rate (AR) for a study and the individual adjudication agreement rate (AAR) for each reader. The RDI indicates the percentage of disagreed cases for a given reader across the total number of cases read, as defined in Equation (1) where a low RDI value indicates better reader performance and high RDI value indicates poor reader performance.

$$\text{RDI} = \frac{\#\text{ of cases where adjudicator disagreed with given reader}}{\text{Total \# of all cases read}} \qquad (1)$$

RDI uniquely combines aspects of both AR and AAR to its advantage. RDI considers the subjects for which adjudicators disagree with the other reader (as shown in numerator of Equation 1). RDI also considers adjudicator disagreement relative to total number of cases read (as shown in denominator for Equation 1) [11-13]. For a specific trial, RDI is best presented along with the reviewers' mean ± standard deviation (SD) value.
The aim of our current work is to investigate use of RDI as an objective measure of reader fatigue. It is important to gain insight whether a reviewer's performance is affected by undertaking single versus multiple type of studies.

From our development of RDI methods, M. Sharma et al. [11] reported in ASCO 2018 that RDI could identify all discordant readers in a retrospective analysis of twenty studies, whereas AR and AAR identified the discordant reader in only 13 and 12 of the twenty studies, respectively. RDI was found to be a quality indicator by effectively combining AR and AAR in identifying those outliers. In the 2019 presentation to SPIE conference on an investigation involving three other oncology trials, M. Sharma et al.[12] again reported RDI as being more reliable index than AR and AAR. In short, the authors have already shown effectiveness of RDI is an excellent indicator for identifying discordant readers in clinical trials.

The next step to be taken was to further expand use of RDI. A retrospective review of BICR adjudication data of twelve oncology clinical trials was carried out. These trials used several established imaging end-point assessment criteria, including: RECIST (Response Evaluation Criteria in Solid Tumors), the Lugano classification and iwCLL (International Workshop on Chronic Lymphocytic Leukemia). RDI for each reviewer was used to identify the discordant reader when only approximately 10% of the total reads were completed for each study and compared with AR and AAR on an ongoing basis throughout the study. As predicted, RDI consistently identified the most discordant reader consistently across all twelve studies, while AR & AAR did not. The new understanding was that RDI, when calculated as early as the 10% of total reviewed cases, demonstrated a positive predictive value (PPV) of 91% and negative predictive value (NPV) of 93% [13].

M. Sharma et al. in the SPIE 2020 conference presented retrospective analysis of readers' RDIs in nineteen different clinical trials. Ninety-two reader performances were examined at five intervals in each trial by forty-three board-certified radiologist readers using established imaging assessment trial criteria. The purpose was to see how well RDI performance above a threshold at progressive monitoring intervals would "flag" a potential overall end-point performance "issue" for that specific reader. The prediction of exceeding threshold i.e. one standard deviation above a study mean RDI was done. Sensitivity, specificity, PPV and NPV were determined for the predicted performance outcomes. Interpretation of multiple "flags" for each trial to improve the aforementioned metrics was explored. Multiple flags enabled statistically improved specificity and PPV [14].

2. METHODS

Data Acquisition: A retrospective analysis of reviewers' RDIs in four different clinical trials were performed to evaluate the reader fatigue using RECIST 1.1 assessments in a BICR. Fourteen reviewer performance was examined using 3750 subjects with 15105 total timepoints in the trials. These individual trial reviews were conducted by 14 board-certified radiologist reviewers using several established imaging assessment trial criteria. Data were blinded with respect to study names, study sponsors, treatment arm, subjects and reviewers. Similar study designs were selected to focus on analyzing RECIST 1.1 criteria assessments discordance rates. Reader fatigue was correlated with RDI. RDI was calculated against each timepoint. RDI Timepoints (RDI_TP) was plotted against timepoint (TP) volume in various ways like per study, across all studies and also %volume of the study in question compared to all studies data.

Table 1 mentions the details of the four studies with number of subjects, number of timepoints, number of readers and time period in days. The highest number of subjects (1767) were in study 3 whereas least number of subjects (505) were in study 1. Number of timepoints ranged from 2901 to 4916. The number of readers ranged from 4 to 7. The time period in days for each study was 181 days. Total number of subjects were 3750 and total number of timepoints were 15105.

Table 1: Details of the four studies including number of subjects, timepoints, readers and time period (in days).

Studies	Number of Subjects	Number of Timepoints	Number of Readers	Time Period(days)
1	505	4916	6	181
2	614	3754	7	181
3	1767	3534	5	181
4	864	2901	4	181
Total	**3750**	**15105**		

Table 2 mentions details about the 14 readers with respect to number of studies, number of subjects and number of timepoints. Number of studies varied from 1 to 3 with number of subjects varying from 310 to 783. The number of timepoints varied from 930 to 3459 for different readers.

Table 2: Details about the number of studies, subjects, timepoints with respect to different reader

Reader	Number of Studies	Number of Subjects	Number of Timepoints
1	3	783	3459
2	2	442	1307
3	2	310	930
4	3	400	1603
5	2	494	1890
6	2	386	1140
Others (8)	1 each	1406	4761

Data Analysis Methods: After obtaining the data of four clinical trials with high volume reads, data prepared for the analysis using R programming script (The R Foundation for Statistical Computing Platform, version 4.1.0 (2021-05-18) - "Camp Pontanezen") augmented by RStudio: Delaware Public Benefit Corporation (PBC) and a Certified B Corporation®.

3. RESULTS

For each clinical trial/ study, study specific read data was collected from database. The following information was collected as shown in Table 3:

- Total number of subjects and timepoints read by each reader for the duration of 6 months from January 2021 to June 2021.
- Adjudication Agreement rate at the Timepoint level was included.
- Reader Disagreement Index (RDI), the improved indicator for monitoring performance, was calculated as reader performance metric. We consider RDI as a fundamental measure of reader performance for better clinical outcomes.

Table 3 shows reader-wise data for the four studies. The details include timepoints, TP AAR, TP RDI%, TP All studies and %TP versus TP All. The number of subjects varied from 57 to 369 and timepoints varied from 242 to 1956 for different readers. TP AAR ranged from 4 to 69 and TP RDI% ranged from 5 to 26. The %TP versus TP All was lowest at 6% for reader 1,2,3 and other readers and highest at 44% for other readers in study 3.

Table 3: Reader wise data for the four high volume studies

READER	STUDY	#SUBJECTS	#TIMEPOINTS	TPs_AAR	TPs_RDI%	TPs_ALL studies	%TPvsTPs_All
Reader 1	1	317	1956	58	10	12994	15%
	2	186	785	63	9	12994	6%
	3	280	718	57	17	12994	6%
Reader 2	1	73	367	39	19	6265	6%

	4	369	940	33	17	6265	15%
Reader 3	2	57	242	4	26	4300	6%
	4	253	688	49	14	4300	16%
Reader 4	1	95	576	58	14	4868	12%
	2	120	487	62	5	4868	10%
	4	185	540	69	8	4868	11%
Reader 5	1	184	1129	51	9	8168	14%
	4	310	761	57	10	8168	9%
Reader 6	2	116	475	69	7	1736	27%
	3	270	665	51	23	2622	25%
Other Readers	1	76	413	42	16	5578	7%
	3	276	690	44	25	2129	32%
	3	319	805	48	24	1819	44%
	2	174	697	35	16	3919	18%
	1	73	432	24	21	4259	10%
	2	158	741	50	12	9069	8%
	3	251	656	53	19	2034	32%
	2	79	327	59	5	5074	6%

Table 4 shows study-wise reader participation at timepoints for four studies. From the table 4, it can be observed that there were significant number of trials where the readers were participating during the same time and was represented as the Timepoints read by the readers across all studies and the percentage of timepoints in the current study across all timepoints. The number of subjects varied from 57 to 369 and timepoints varied from 242 to 1956 for different readers. TP AAR ranged from 4 to 69 and TP RDI% ranged from 5 to 26. The %TP versus TP All was lowest at 6% and highest at 44%.

Table 4: Trials describing reader participation at timepoints

STUDY	READER	#SUBJECTS	#TIMEPOINTS	TPs_AAR	TPs_RDI%	TPs_ALL studies	%TPvsTPs_All
1	Reader 1	317	1956	58	10	12994	15%
	Reader 2	73	367	39	19	6265	6%
	Other	76	413	42	16	5578	7%
	Reader 4	95	576	58	14	4868	12%
	Reader 5	184	1129	51	9	8168	14%
	Other	73	432	24	21	4259	10%
2	Reader 1	186	785	63	9	12994	6%
	Reader 3	57	242	4	26	4300	6%
	Reader 4	120	487	62	5	4868	10%
	Other	174	697	35	16	3919	18%
	Other	158	741	50	12	9069	8%

	Other	79	327	59	5	5074	6%
	Reader 6	116	475	69	7	2622	18%
3	Reader 1	280	718	57	17	12994	6%
	Other	276	690	44	25	2129	32%
	Other	319	805	48	24	1819	44%
	Other	251	656	53	19	2034	32%
	Reader 6	270	665	51	23	2622	25%
4	Reader 2	369	940	33	17	6265	15%
	Reader 3	253	688	49	14	4300	16%
	Reader 4	185	540	69	8	4868	11%
	Reader 5	310	761	57	10	8168	9%

Figure 1: Plot of timepoints (in a particular study) versus RDI TP.
Adj R2 = 0.07 and P < 0.05 with grey area around the slope by default, being the 95% confidence level interval for predictions from a linear model.

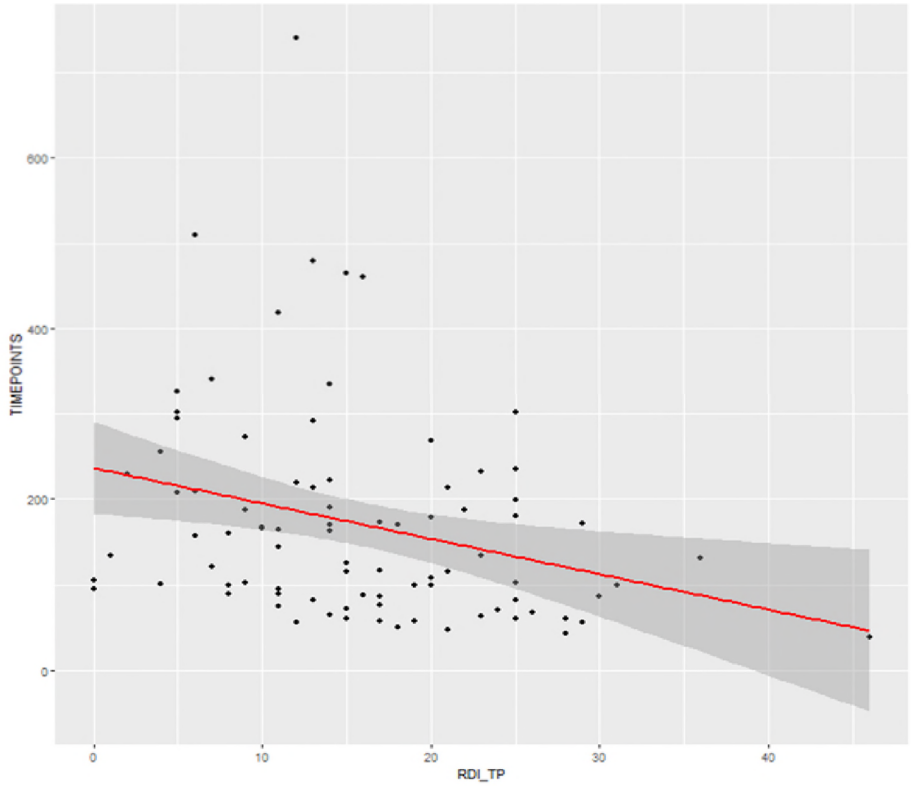
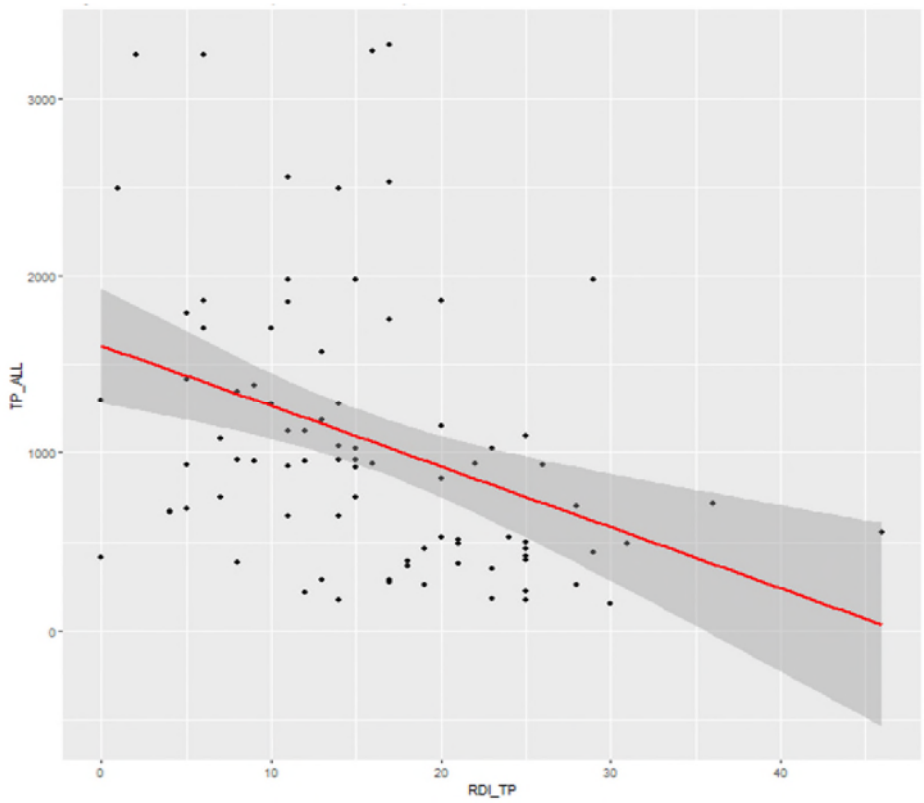


Figure 2: Plot of TP All (across all studies) versus RDI TP.

Adj R2 = 0.13 and P < 0.05 with grey area around the slope by default, being the 95% confidence level interval for predictions from a linear model.



RDI is proven to be an effective indicator for monitoring central reader performance and identifying discordant reader. We realized this could also be a great surrogate to monitor on an ongoing basis against reader volume, which could in turn be a good surrogate for reader fatigue. Given that ocular movement monitoring devices-based measurement of reader fatigue is not possible in clinical trial setting, RDI may be used on an effective daily / weekly / monthly basis against read volume for monitoring reader fatigue which affects variability and thus read quality. Increasing RDI trend may be used as an early indicator for potential reader quality and fatigue if a particular reader is concurrently reading on multiple trials.

4. DISCUSSION & CONCLUSIONS

Based on our previous presentations at various conferences including SPIE 2019 and SPIE 2020, RDI has become novel measure of reader variability assessment and proven to be a useful indicator for monitoring read quality in clinical trials by identifying a potentially discordant reader. Given that the trials we evaluated were already actively monitored for quality, we could not identify obvious poor performers by RDI which could then be correlated with obvious signs of compromised read quality and / or fatigue. Though, statistically not significant an early trend towards low RDI with higher volume suggests better reader recall value for clinical trial rules.

Thus, more extensive study on multiple trials may be carried out to understand the role of RDI as a tool for prediction of reader fatigue in prospective setting correlating the number of trials a reader reads on in parallel, given that poor performance in regulatory setting is uncommon by itself.

REFERENCES

[1] Stec Nadia, Arje Danielle et al. A Systematic Review of Fatigue in Radiology: Is It a Problem? *AJR* 2018; 210:799– 806.

[2] Taylor-Phillips S, Stinton C. Fatigue in radiology: a fertile area for future research. *Br J Radiol* 2019; 92: 20190043.

[3] Waite Stephen, Kolla Srinivas et al. Tired in the Reading Room: The Influence of Fatigue in Radiology. *J Am Coll Radiol* 2017; 14(2): 191-19. http://dx.doi.org/10.1016/j.jacr.2016.10.009.

[4] Krupinski Elizabeth, Reiner Bruce I. Real-Time Occupational Stress and Fatigue Measurement in Medical Imaging Practice. *J Digit Imaging* 2012; 25:319–324. DOI 10.1007/s10278-011-9439-1.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] Cohen KL, Gönen M, Ford RR. Monitoring Reader Metrics in Blinded Independent Central Review of Oncology Studies. *J. Clin Trials* 2015;5:4. doi:10.4172/2167-0870.1000230.

[10] Ford RR, O'Neal M, Moskowitz SC, Fraunberger J. Adjudication Rates between Readers in Blinded Independent Central Review of Oncology Studies. *J Clin Trials* 2016; 6-5.

[11] Sharma Manish, Singareddy Anitha, O'Connor Michael, Fotinos-Hoyer A. Kassel, Karve Sayali, Enus Nicholas, Clark Daniel. Reader disagreement index (RDI) as an indicator of reader performance. *Journal of Clinical Oncology* 2018; 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

[12] Sharma Manish, O'Connor Michael J, Singareddy Anitha, 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.

[13] Sharma Manish, Bohnsack Oliver, O'Connor Michael, Shao Yibin, Enus Nicholas, Karve Sayali, Fotinos-Hoyer A. Kassel. RDI as a method for reviewer performance monitoring in BICR setup for improving data quality. *Journal of Clinical Oncology* 2019; 37, no. 15_suppl, Published online May 26, 2019; DOI: 10.1200/JCO.2019.37.15_suppl.e18082.

[14] Sharma Manish, O'Connor Michael J, Singareddy Anitha, Madasu Madhuri, Enus Nicholas, Fotinos-Hoyer Kassel, Shao Yibin. Using "Reader Disagreement Index" as a predictive reviewer performance monitoring tool for timely intervention. *Proc. SPIE 11316*, Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment, 113160A (16 March 2020); doi: 10.1117/12.2549980.