Therapeutic Innovation & Regulatory Science (2021) 55:1111–1121
https://doi.org/10.1007/s43441-021-00316-6

DIA

REVIEW

Radiologists and Clinical Trials: Part 1 The Truth About Reader Disagreements

Annette M. Schmid, PhD[1] · David L. Raunig, PhD[1] · Colin G. Miller, PhD[2] · Richard C. Walovitch, PhD[3] · Robert W. Ford, MD[4] · Michael O'Connor, PhD[5] · Guenther Brueggenwerth, MD[6] · Josy Breuer, MD, PhD[7] · Liz Kuney, MS[8] · Robert R. Ford, MD[9]

Received: 11 October 2020 / Accepted: 18 June 2021 / Published online: 6 July 2021

Abstract

The debate over human visual perception and how medical images should be interpreted have persisted since X-rays were the only imaging technique available. Concerns over rates of disagreement between expert image readers are associated with much of the clinical research and at times driven by the belief that any image endpoint variability is problematic. The deeper understanding of the reasons, value, and risk of disagreement are somewhat siloed, leading, at times, to costly and risky approaches, especially in clinical trials. Although artificial intelligence promises some relief from mistakes, its routine application for assessing tumors within cancer trials is still an aspiration. Our consortium of international experts in medical imaging for drug development research, the Pharma Imaging Network for Therapeutics and Diagnostics (PINTAD), tapped the collective knowledge of its members to ground expectations, summarize common reasons for reader discordance, identify what factors can be controlled and which actions are likely to be effective in reducing discordance. Reinforced by an exhaustive literature review, our work defines the forces that shape reader variability. This review article aims to produce a singular authoritative resource outlining reader performance's practical realities within cancer trials, whether they occur within a clinical or an independent central review.

Keywords Visual perception · Independent review · Reader disagreement · Image interpretation · Radiology · Clinical trials

Introduction

Today's clinical trials for cancer therapies regularly depend upon the interpretation of medical imaging to establish efficacy. Of the 20,682 cancer trials registered in clinicaltrial.gov over the last decade,[1] 67% feature endpoints hinging upon medical imaging.[2] Image readers for cancer trials, typically radiologists or nuclear medicine experts, can be on-site in the clinical setting or comprise teams of selected experts performing an off-site Blinded Independent Central Review (BICR). The BICR serves as the backdrop for this paper.

David Raunig and Annette Schmid contributed equally to the development and writing of the manuscript. David Raunig contributed to this article while employed at eResearch Technology.

✉ David L. Raunig
draunig@snet.net

1 Takeda, 300 Massachusetts Ave, Cambridge, MA 02139, USA

2 The Bracken Group, Newtown, PA, USA

3 WorldCare Clinical, Boston, MA, USA

4 Thomas Jefferson University Hospital, Philadelphia, PA, USA

5 Parexel, Billerica, MA, USA

6 Bayer AG, Berlin, Germany

7 Consultant, Lutherstadt Wittenberg, Saxony-Anhalt, Germany

8 Kuney Consulting, LLC, Syracuse, NY, USA

9 Bioclinica, LLC, Princeton, NJ, USA

[1] clinicaltrials.gov search terms: Recruiting, Active, not recruiting, Completed, Suspended, Withdrawn Studies | Interventional Studies | Oncology | Phase Early Phase 1, 1, 2, 3 | Start date from 01/01/2011 to 12/31/2020).

[2] clinicaltrials.gov search terms: imaging OR recist OR Lugano OR Cheson OR mri OR ct OR ultrasound OR pet OR rano OR Choi OR PFS OR ORR OR BOR | Recruiting, Active, not recruiting, Completed, Suspended, Withdrawn Studies | Interventional Studies | Oncology | Phase Early Phase 1, 1, 2, 3 | Start date from 01/01/2011 to 12/31/2020.

The purpose of the BICR is to provide an image assessment independent of bias and potential functional unblinding that may exist when patient health information is known. The FDA's current guidance for industry affirms that the BICR "enhances the credibility" and "better ensures the consistency" of imaging assessment [1]. The BICR is employed for most but not all cancer trials of all phases but deemed specifically important in Phases 2 and 3 and necessary in single-arm or unblinded studies. The most common BICR design is referred to as "2 + 1," in which two independent readers assess the same images for each study subject, and one adjudicator settles endpoint disparities between the two outcomes. The duplication of two readers with the addition of the adjudicator enhances the reliability of endpoint assessment and makes the measurement of reader performance possible. Though site readers are expected to have the same sources of variability, measurement of that variability is most often not practical if even possible. However, any two readers will inevitably differ in their interpretation to some degree. Research over the last 70 years shows that that disagreement between two radiologists is remarkably consistent throughout the decades and also comparable to the inherent variability between two physicians within any field. Necessarily, the variability that leads to disagreement is present in all radiologists, whether on-site or at a central reading facility. While differences in interpretation do not inevitably nor inherently equate to error, understanding the sources of these differences and associated significance will help guide appropriate action to minimize and control these differences when they can be controlled and understand them when they cannot. The question of whether an independent read using multiple readers provides greater benefit than using individual site investigators is discussed in detail elsewhere [2–7]. Methods to measure and monitor for these differences and their relationship with sources of variability are presented in the companion paper to this manuscript [8].

Reader Disagreements are Consistent with Diagnostic Disagreements in Other Areas of Medicine

The breadth of literature supports the broader medical community's view that disagreement between experts is inevitable and at times necessary in all fields of medicine, including radiology and nuclear medicine. The Society to Improve Diagnosis in Medicine [9]. was established in 2011 to improve misdiagnoses in the clinic, including those made by radiologists. The primary concern over misdiagnosis is the main motivation behind 70 years of exhaustive focus into the causes of perceptual error in imaging [10]. Contemporary medical imaging associations, such as the Medical Image Perception Society (MIPS), have made extensive contributions over many years to the study of medical image interpretation [11]. Since BICR evaluations include a diagnostic component of lesion assessment, as well as evaluating disease status over time, the variability between independent clinical trial readers will also necessarily include some of the challenges seen in the clinic [12–20].

Disagreement among physicians is an integral part of medicine. Interestingly, disagreement rates are remarkably consistent across the decades for different types of image evaluations, across medical specialties and different technologies, falling between 20 and 40% [20–35] (see Table 1 for a summarized sample of research on evaluator agreement)

Earlier research on radiological discordance focused on disagreements in diagnoses. Other research specifically dedicated to radiological image perception concluded that radiological disagreements are expected for radiologists as they are for any physician [36–44]. In a landmark article on radiological error from 1959, Henry [45] stated

> Even experienced physicians are found to have a measurable degree of 'observer error' due apparently to the so-called human equation… In evaluating pairs of serial roentgenograms, (two physicians are) apt to disagree … in about one-third of the cases and with (themselves) in one-fifth of them.

Remarkably, even when considering technological advancements in scanners, little has changed in the rate of reader discordance over time. Radiological disagreement rates of approximately 30–40% reported by several papers since 1959 are consistent with the rates shown by Ford in 2016 across a variety of oncologic indications [45–47] [, holding steady across different response criteria, whether quantitative or subjective in form [26, 35, 45–49]. This consistency may be due to only 5–10% of the information for visual perception coming from the retina while 90–95% comes from different regions of the brain including the cortex and brain stem [50]. Therefore, the majority of the inputs that affect visual perception are resident in the brain at the time the images are evaluated.

Why Expert Readers Disagree

The key to taking appropriate actions to minimize reader disagreement in a clinical trial setting is understanding what sources of variability are involved. Assessing medical images demands cognitive tasks such as reasoning, problem-solving, and visual perception. Within these tasks, clinical trial readers must not only identify and determine the state of the disease but also when the disease changed enough to cross a criteria threshold. The greatest contributing factor of inter-reader variability originates from a radiologist's own

Table 1 A Sample of Evaluator Agreements for Different Specialties Since 1947.

Author	Year	Disagreement or Error Rate	Type of Assessment
Birkelo et al.	1947	Inter-reader: 35% Intra-reader: 20%	5 radiologists Tuberculosis radiological diagnosis Film
Thiesse et al.	1997	Major disagreements—40% Reasons: tumor measurements, selection of measurable targets, intercurrent diseases, and radiologic technical problems	Renal cell carcinoma Disagreement with committee of tumor response WHO criteria basis
Gwyther et al.	1997	Disagreement with response: 39%	Epithelial ovarian cancer Response: WHO criteria 2 Independent readers
Rubenfeld et al.	1999	$\kappa = 0.55$ 32% had ≥ 5 dissenters	Acute respiratory distress syndrome diagnosis (CT) 21 experts $N = 28$
Wormanns et al.	2000	5 mm slice thickness Detection disagreement: 38% Size category: $\kappa = 0.61$	Pulmonary nodules 2 readers Detection and size 23 patients 286 nodules
Aldape et al.	2000	Disagreement: 23%	Glioma Digital pathology/neuropathology Diagnosis $N = 457$
Pandolfino et al.	2002	Intra-observer κ Experts: 0.55 Trainees: 0.44 Inter-observer κ Experts: 0.56 Trainees: 0.46	Endoscopic scoring of esophagitis Experts and trainees
Scholten et al.	2004	FIGO disagreement: 30% ($\kappa = 0.4$)	Endometrial carcinoma Digital Histology (FIGO) $N = 800$ 2 independent pathologists
Gietema et al.	2006	Discrepant volumes: 10.9% Inter-reader Spearman Correlation: r = 0.99	Lung cancer $N = 232$ nodule detection ($n = 430$) Local and Central reader Volume
Hricak H, et al.	2007	Staging CT: $\kappa = .26$ MRI: $\kappa = .44$ Visualization CT: $\kappa = .16$ MRI: $\kappa = .32$ Sens/Spec Sens: CT = .26 MRI = .48 Spec: CT = .92 MRI = .79	Cervical cancer Diagnosis $N = 326$ 4 radiologists (CT) 4 radiologists (MRI)
Hersh et al.	2007	All combinations of readers Disagreement = 25%	Lobe- predominant emphysema HRCT $N = 30$ Pulmonologists and radiologists
Suzuki et al.	2010	Inter-reader agreement $\kappa = 0.53$ (95% CI 0.33-0.72) Intra-reader agreement $\kappa = 0.86$ (95% CI 0.76-0.96)	Breast and Colorectal cancer $N = 39$ RECIST response 2 radiologists
Ibrahim et al.	2011	Inter-reader agreement for subarachnoid hemorrhage $\kappa = 0.41$ (95% CI 0.33–0.49)	Aneurysmal subarachnoid hemorrhage $N = 413$ 1 neurosurgeon 1 neuroradiologist

expertise in applying the essentially subjective aspects of the response criteria.

Factors that affect reader performance can be roughly defined as controllable, (e.g. experience/expertise, fatigue, and environment), less-controllable (e.g. daily disposition, stress, and internal biases), and not controllable (e.g. random measurement variability and biological heterogeneity). The less controllable factors comprise 89% of radiological disagreement [31, 51]. Factors such as experience level and reader fatigue can be controlled to some degree and are recognized by regulatory authorities as factors to plan for and monitor and are important enough to be addressed in the FDA Guidance for Industry [1].

Controllable Factors in Reader Performance

Aside from the ambiguity inherent in imaging complex anatomy, sources of reader variability that can be controlled include expertise, training, the reading environment, and setting, including the risk of fatigue [52–54]. Radiologists and other readers each have unique levels and types of training and experience that can also contribute to discordance and can be controlled to an extent by the choice of readers for the study. Familiarity with specific disease indications and clinical trial review, the extent of specific reader training and knowledge, and familiarity with the response criteria can also sway interpretation. Moreover, controllable factors unrelated to the individual may also have a considerable impact. These can include the imaging technique's quality and limitations, the number of response categories, tumor growth rates relative to image sampling rates, and specific tumor feature characteristics.

Here, we provide additional detail to the major discordance categories to explain why they occur to lay the groundwork for reader performance monitoring methods described by Raunig et al. [8] in a companion paper.

Image Interpretation and Experience

Reading medical images requires the detection, interpretation, and appropriate labeling of visible information of interest. Visual information that includes complex shape, texture, and intensity of the entire image is processed by the visual pathways in a manner that is strongly influenced by experience and higher functions including learning and memory [41, 55, 56]. Borradaile et al. authored a review of 40 oncology clinical trials across 12 different indications with 12,299 participants and concluded that differences in expert visual interpretation commonly referred to as "medical image perception," comprised 77% of the disagreements [57]. Their figure is remarkably consistent with the estimate of 80% independently reported by Kim and Mansfield for radiological diagnostic errors [58].

The importance of clinical trial experience in addition to clinical experience has been noted in the industry as exemplified by the advertisement of the Massachusetts General Hospital of their central reading services:

> With over 20 years of clinical trial experience, our radiologists understand the unique needs of CROs [Contract Research Organizations] and pharmaceutical companies and are well equipped to handle even the most challenging of trials [59].

Radiological experience plays a particularly critical role when new findings may represent benign or unrelated conditions mistaken for new metastasis or disease detected in less common manifestations or locations. For example, a pulmonary embolus can appear to be a new lung lesion, impersonating new pulmonary metastasis. Experience and training on the specific implementation of the clinical trial criteria are also critical. The proliferation of response criteria, over 20 in oncology, increases the chance that the readers, site or central, will misinterpret the criteria and, therefore, commit the same procedural error. For example, in a trial involving prostate cancer and the newly released PCWG 3 criteria [60], several readers evaluated according to the older criteria, PCWG2 [61]. The result of the errors was that the scans were re-opened and required re-reading, the readers were required to undergo refresher training, and a plan of corrective actions and preventive actions was created and implemented including a diary entry into the trial master file.

Though there is widespread agreement that more experienced radiologists have better diagnostic sensitivity and specificity than less experienced radiologists, there is no defined threshold for the number of years' experience needed to successfully read in a clinical trial though. Some research indicates that between 5 and 10 years of experience as a practicing radiologist may be a useful guideline for recruiting candidate readers [62–64]. Additionally, Tucker et al. reported that fewer than 80,000 cases read was an apparent threshold for decreased diagnostic accuracy. A search of clinicaltrials.gov for Phase 2 and 3 studies using RECIST for PFS resulted in 3429 studies over the last 10 years for an average of 175 subjects/trial which may be used to approximate the number of clinical trial cases read when the reader curriculum vitae indicates only the number of clinical trials experience and not the total number of cases.[3] Interestingly, the interaction of reader experience and fatigue having a greater influence on performance for readers with less experience [29, 30, 47]. However, this may not always be true since, at times, information on newer scanning techniques

[3] (Search terms: RECIST | Recruiting, Active, not recruiting, Completed Studies | Interventional Studies | Cancer | PFS | Phase 2, 3 | Start date from 01/01/2011 to 12/31/2020).

that were previously unavailable may compensate for a reader's lack of experience [65].

The following are recommended as considerations when choosing either blinded independent readers for central reads or sites and radiology departments when using site readers:

- Experience
    - Clinical experience in radiological or nuclear medicine evaluations of the specific indication measured in both years as a practicing clinician and number of patients evaluated;
    - Criteria experience or training
    - Experience with reading for a clinical trial, including the phase of the trial which may indicate experience with timelines, criteria, and reader workload.
- Fatigue
    - The numbers of readers used in a pool to offset the workload on the readers at different parts of the clinical trial (e.g. interim analyses) and at the end of the study.
    - Monitoring or restricting the number of cases read in a single read session or over a longer period (e.g. week or month).

Selecting and Measuring Target Lesions

For most metastatic cancers, response criteria are generally concerned with measuring the change in the patient's tumor burden. However, measuring all visible lesions in all patients is simply impractical. Therefore, following only a sampled subset of lesions over time is the basis behind most objective response criteria. Radiologists must be able to select 'relevant" lesions at baseline that they believe represent the disease burden of the patient and that will continue to be accurately measurable throughout treatment. A radiologist's ability to select suitable target lesions is also dependent on their interpretation of what constitutes a suitable target lesion based on experience. For example, a lesion that meets measurability criteria may also later coalesce on subsequent timepoints, i.e. hilar or mediastinal lymphadenopathy typically leading to changes in measurements inconsistent with the change in the disease state. In these cases, a radiologist who is experienced in reading for a clinical trial may be more likely to choose that lesion as non-target at baseline. Sridhara, et al. point out that a target lesion that cannot be followed by at least one reader can result in missing data and a not evaluable assessment [66]. Examples of this, that members of PINTAD have observed, occurred in several studies when the site chose the target lesions for the central readers.

Complicating matters, the specific target lesions readers select often differ especially in patients with numerous lesions. As the percentage of change will vary and meet certain thresholds at different times depending on the set of lesions selected these differences in selection have been identified as a major reason for reader disagreement [24, 67]. Nevertheless, studies show that allowing readers to independently select target lesions does not affect the overall study result and increases the overall reliability of the result by reducing sampling error of the target lesions that might occur by leaving the target lesion response up to a single reader. [68, 69].

Detecting New Lesions

New lesions that are still small may miss detection or, even if detected, reader comments in many clinical trials indicate that they want to wait to confirm that it is a growing lesion and not a non-malignant finding. Disagreements on whether a "new lesion" existed at baseline for disease-free survival endpoints can lead to the casewise exclusion of that patient for analysis. Interestingly, researchers from MIPS point out that radiologists can miss or misdiagnose lesions even when directed to the location of interest [11].

The detection of new lesions is not only a matter of perception but also of signal versus noise—small lesions in inherently noisy images. To help increase signal, supplemental imaging modalities may also assist the readers when specified or designed to be acquired [70, 71]. Errors by a single reader in new lesion detection account for approximately 10% of all discordances [57].

Recognizing Non-target Lesion Progression

Unequivocal non-target progression should reflect growth in which the "overall tumor burden has increased sufficiently to merit discontinuation of therapy" [72]. Accordingly, disagreements between independent readers regarding non-target lesion progression occur [73–77]. Disagreement on non-target lesion progression comprises about 10% of all disagreements [78] and, while this constitutes a small percentage of all disagreements, perceptual disagreements can be a source of controversy when discussing patient care. Objective evaluation of non-target lesion response to treatment is chiefly dependent on noticeable morphological degrees of growth, reader experience, and the readers' internal thresholds of when to call unequivocal tumor progression (see also reader bias below). The likelihood and degree of this kind of disagreement can be reduced by ensuring that all readers are jointly trained in the indication, the modality, the criteria, and, most importantly, covering the scenarios that constitute unequivocal progression.

Lesion Measurement

Well-defined, oval, or round lesions with clear lesion-to-background discrimination are easier to measure and result in less variability than complex lesions with diffuse or spiculated borders or those with poor discrimination from the background [79, 80]. Some indications, such as hepatocellular carcinoma or ovarian cancers, are particularly challenging to measure. A lesions' shape, conspicuity, or diffuse or infiltrating borders, require varying degrees of subjectivity in their measurements [46]. Lesions are also subject to slight movements and deformations due to patient positioning, breathing, swallowing; furthermore, the same lesion can demonstrate changes in its diameter, even upon immediate re-imaging [79].

Also, tumor size is typically measured in a single plane by the longest and/or shortest (i.e. widest) diameter, depending on the lesion type and criteria. RECIST measures solid, non-lymphatic tumors along the longest diameter. Therefore, measurement differences among readers due to differences in the determination of the measured edge or the longest axis can affect the target lesion response. [81] Measurement variability between readers for the same lesion, for most tumor types, can be considered to be only a minor contributor to overall reader variability (intraclass correlation coefficient = 0.991) [82].

Measurement variability can and does lead to different response categories across readers. One example seen in a clinical trial was an assessment of stable disease based on a 19.7% increase in the sum of diameters and the other reader assessing progressive disease based on a 20.1% increase in a single brain metastatic tumor. The adjudicator chose stable disease, which resulted in no progression event at that visit. The actual percentage change was small but the response category difference highlighted disagreements that are inevitable with criteria that rely on thresholds.

Image Quality

Inherent in any reader assessment accuracy is the quality of the image, itself, tempered by one's ability to perceive the true disease state from poorly acquired images. Alpert and Hillman reported that from 10 to 24% of diagnostic errors in the clinic are associated with low image quality [83]. Certainly, this would also apply to the central review. Low-quality bone scans, computed tomography (CT) scans with motion artifacts, differences in contrast enhancement, inadequate contrast levels, or incomplete data are all image quality issues that make reliable assessments difficult and can increase reader differences. These nuances demand consummate and careful focus. Clear guidance on the expected image acquisitions needs to be provided and the consequences of adjustments due to oversight or for example patient condition understood (e.g. patient cannot tolerate a full dose of intravenous contrast) by trialists and trial sponsors. Pre-study training should prospectively discuss these scenarios, in particular in the context of implications for the assessment criteria (e.g., the timing of contrast in mRECIST for HCC).

Missing Clinical Information

Clinical information may direct the visual search to specific anatomical locations, or clinical information can provide context on a particular finding's nature. Unlike in clinical practice, where the practice of medicine integrates objective disease progression with patient-related medical care factors, such as toxicity, medications, incidental findings, and overall clinical health, objective assessment of a patient's response to treatment by imaging is fundamentally based on the reader having no information on the patient that is not pre-determined as part of the reader assessment. Accordingly, patient-reported health status, typically available to the clinical radiologist, is not available to the clinical trial reader to ensure an objective assessment and reduce the possibility of biases to influence the assessment. For example, a reader who knows that the patient's deteriorating health status may be biased to confirm that knowledge by assessing progressive disease (i.e. confirmation bias). The reverse of confirmation bias, anchoring bias, might occur if the reader is biased by the patient's health status and then fails to adjust their assessment in light of contradictory radiological information. A complete list of the 10 biases that radiologists are prone to was compiled by Busby et al. [54]. To mitigate the risk of bias and to also include clinical information critical to the assessment (e.g. biopsy), a careful and prospective determination of what clinical data is helpful in the context of the disease and how it is to be integrated with the criteria is strongly recommended and should be included into the image review charter.

Discussion

In 1996, the Clinton-Kessler Oncology Initiative accelerated cancer drug development. With this, centralized imaging rose in importance, becoming a prominent fixture in oncology clinical trials. Accordingly, imaging core labs set out to establish specific read processes for clinical trials. Under the scrutiny of both regulators and sponsors, these processes evolved to establish controls to ensure the reliability of blinded assessments by measuring and monitoring central reader performance, specifically by disagreements between paired readers on teams of two. Though medical imaging is common to most oncology trials, the familiarity of study personnel with what constitutes reasonable disagreement

and when or what corrective measures should be taken can vary greatly. We present in this paper a review of the literature as well as the experience of PINTAD members on why readers disagree and what kind of disagreement is to be expected. Fortunately, while errors are bound to occur when multiple readers read a case the errors are almost always limited to one of the readers since the chance that multiple readers committing the same error is small. Therefore, multiple reader paradigms allow for both an overall measure of disagreement and methods to eventually identify which of the readers contributed to that disagreement, provided in great detail in the companion to this paper [8].

A trend seen both by PINTAD members and within the literature on ICL reader disagreements is an increase in the desire to reduce reader disagreement using methods that also reduce the independence of the two readers. One example is to have a single reader choose the target lesions. This particular read design places target lesion response almost entirely on the abilities of the lesion-selection reader and reduces the adjudication of disagreement to a choice of how the readers measure lesions and not on the best set of target lesions. The common use of computer-aided measuring tools would reduce target lesion disagreement to near-zero but at the risk of a less-than-optimal choice of target lesions. Another example of misplaced efforts to reduce reader disagreement allows the readers to discuss the cases before the assessment which constitutes, essentially, a consensus assessment and which is specifically not accepted by the regulatory authorities as a multiple reader assessment. These and other examples demonstrate the risk of methods that, in essence, force the readers to agree without regarding the statistical impact on endpoint accuracy.

Many factors that shape reader variability are common to both on-site and off-site reviews. However, literature comparing imaging core lab and site reads is incomplete. The seminal meta-analysis of 27 studies that concluded equivalence between central and site readers was limited by over half of the studies having sites and core labs communicate with each other (i.e. not independent) or having protocol amendments that required mitigation of site-related bias [84–87]. Nevertheless, the processes unique to each setting introduce different degrees of control for the reader variabilities. The largest and most impactful differences include (1) the central read's consistent use of two independent radiologists, and (2) the ability of the central review process to monitor readers for both short-term and long-term trends. In most cases, monitoring site readers for bias, drift, and errors is typically impractical if even possible, and periodic retraining for all site readers may be needed in long follow-up studies to help ensure the reliability of the site-assessed results.

In studies that include the use of both site and central readers in a hybrid model of reader teams, the natural discordance between two radiologists can have an additional impact. For example, when enrollment requires a measurable lesion, which is required by RECIST for an assessment of Partial Response (PR), a disagreement by one reader on the presence of a measurable lesion will preclude that reader from assessing a PR. These incidents have been noted by the authors and other PINTAD members, and the following recommendations are made to mitigate the impact of site versus BICR discordance.

- In alignment with planned endpoints, consider central confirmation of baseline requirements, specifically the requirement for at least one measurable lesion to ensure that response is possible, and in studies that measure relapse for a disease-free survival (DFS) endpoint, have a central confirmation of the absence of disease at baseline.
- Require central confirmation of progression or central adjudication of site-central discordance to reduce the possibility of informative censoring.

Independent from the read setting, when the prescribed disagreement rate is significantly above or below the expected rate, an evaluation should be performed to assess whether the observed rate is justified. In the context of well-trained expert readers working under controlled conditions, a higher disagreement rate may reflect the challenges of the interpretation such as mixed responses due to the specific choice of target lesions, or "borderline cases" that hug the thresholds in slow-growing, or visually ill-defined disease, or simply poor image quality. However, if the investigation into unexpected disagreement rates does not suggest the presence of such justifiable differences in interpretation, then inadequate training, fatigue, or other performance-related factors may be the cause. Understanding which disagreements are justified versus which disagreements can and should be limited and managed greatly adds to the effectiveness of any conclusions and possible remedial actions. Most typically, these remedial actions consist of additional reader training, and, depending on the conclusions drawn, can consist of training as simple as a "Read and Acknowledge," or can be more involved such as a discordant case review remotely or in person. Reviewing discordant cases, though not specifically meant to reduce disagreement rate, does so by resolving the reason for discordance. It may also be helpful to review agreed-upon cases to identify reasons for agreement. In most cases, the review is likely to be most beneficial in identifying any misinterpretations of the criteria which may greatly reduce subsequent disagreements.

In recent years, clinical researchers are looking to artificial intelligence (AI) to support radiologist's reads and reduce reader variability [69–72] in addition to potentially replacing the current assessment criteria. A search of the term "AI" in the 2019 RSNA Annual Meeting program

resulted in 310 tutorials, classes, talks, or posters. Large international research collectives such as PRIMAGE, with access to adequate big data sets, are pursuing deep machine learning that could facilitate personalized imaging biomarkers [88]. We consider these efforts as highly promising, despite earlier research in computer-aided radiology that suggested computer-aided detection had not substantially improved the incidence of human error [44, 89–92], and are looking forward to further validation. For the present, at least, imaging interpretation relies on human expertise; and, reader variability remains an unavoidable reality.

A new concern that has arisen recently and sure to become integral in clinical trials for the future is the presence of intercurrent events (ICE) experienced during the COVID-19 pandemic, particularly those ICEs that may make it necessary for oncology patients to be scanned at different imaging centers or even using different modalities [93]. While the guidance recommends consulting with the FDA for the impact of alternative imaging centers on efficacy endpoints and type I and type II error rates, study sponsors may also want to consider the impact of the pandemic on any studies using local evaluations. In the case of site radiologists being unavailable or overworked, studies should have a discussion about incorporating the consistency and availability of using central readers.

Conclusion: The Significance of Reader Variability in Clinical Trials

It is quite clear that two independent experts will always disagree to some extent. Disagreement rates of 25% to 40% on the interpretation of an image are a reasonable benchmark, based on seven decades of consistent findings. Importantly, variability among readers does not necessarily indicate inadequate performance; instead, it often reflects natural and expected differences in all of its forms and may reveal where multiple interpretations are reasonable. In controlled and monitored reader environments, unexpected levels of disagreement should be flags for further investigation and changes in disagreement rates can be reliable indicators of some type of change in performance. Correctly identifying reasons for reader variability may become even more important in the near future as immune-oncology studies become the standard and different sources of image data noise and confounders become more and more important to mitigate.

The procedures and methods presented in the companion to this article recommend ways to monitor and interpret imaging reviewer performance in most clinical trials [8].

Acknowledgements
The authors would like to express their sincere appreciation to Drs. Anthony Fotenos and Alex Hofling (FDA/CDER) and Dr. Joseph Pierro (eResearch Technology) for their valuable insight, participation, and contributions to the discussions and development of the methodologies within this manuscript. Additionally, the authors express their sincere gratitude to the reviewers for their detailed review and insightful comments.

Author Contributions
AMS and DLR both equally contributed to concept, design, drafting finalizing, and approved all content. CGM contributed to the concept, design, drafting finalizing, and approval of all content. RCW contributed to the concept, drafting and finalizing all content. RWF contributed to the concept, drafting, and finalizing all content. MO contributed to the concept, drafting, and finalizing all content. GB contributed to the concept, drafting, and finalizing all content. JB contributed to the concept, drafting, and finalizing all content. LK contributed to the concept, drafting, and finalizing all content. RRF (Senior author) contributed to concept, design, drafting finalizing, and approved all content. All authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

Funding
No funding sources.

Declarations

Conflict of interest
Annette M. Schmid, David L. Raunig, Colin G. Miller, Richard C. Walovitch, Robert W. Ford, Michael O'Connor, Guenther Brueggenwerth, Josy Breuer, Liz Kuney, Robert R. Ford declare that there are no conflicts of interest.

References

1. FDA. *United States Food and Drug Administration Guidance for Industry: Standards for Clinical Trials Imaging Endpoints*. In: Services UDoHaH, editor. Rockville, MD2018.
2. Eldevik O, Dugstad G, Orrison W, Haughton VJR. The effect of clinical bias on the interpretation of myelography and spinal computed tomography. *Radiology*. 1982;145(1):85–9.
3. Sica GTJR. Bias in research studies. *Radiology*. 2006;238(3):780–9.
4. Ford R, Schwartz L, Dancey J, Dodd L, Eisenhauer E, Gwyther S, et al. Lessons learned from independent central review. *Eur J Cancer*. 2009;45(2):268–74.
5. Amit O, Mannino F, Stone A, Bushnell W, Denne J, Helterbrand J, et al. Blinded independent central review of progression in cancer clinical trials: results from a meta-analysis. *Eur J Cancer*. 2011;47(12):1772–8.
6. Floquet A, Vergote I, Colombo N, Fiane B, Monk BJ, Reinthaller A, et al. Progression-free survival by local investigator versus independent central review: comparative analysis of the AGO-OVAR16 Trial. *Gynecol Oncol*. 2015;136(1):37–42.

7. Wu Y-L, Saijo N, Thongprasert S, Yang J-H, Han B, Margono B, et al. Efficacy according to blind independent central review: post hoc analyses from the phase III, randomized, multicenter, IPASS study of first-line gefitinib versus carboplatin/paclitaxel in Asian patients with EGFR mutation-positive advanced NSCLC. *Lung Cancer*. 2017;104:119–25.
8. Raunig D, Schmid A, Miller CG, Walovitch RC, Noever K, Hristova I, et al. *Radiologists and Clinical Trials: Part 2. Practical Statistical Methods for Understanding and Monitoring Independent Reader Performance Therapeutic Innovation & Regulatory Science*. 2021 Submitted.
9. Medicine StIDi. https://www.improvediagnosis.org.
10. Birkelo CC, Chamberlain WE, Phelps PS, Schools PE, Zacks D, Yerushalmy J. Tuberculosis case finding: a comparison of the effectiveness of various roentgenographic and photofluorographic methods. *J Am Med Assoc*. 1947;133(6):359–66.
11. MIPS. *Medical Image Perception Society 2019*. http://mips.synchrosystems.com/.
12. van den Bent MJ. Interobserver variation of the histopathological diagnosis in clinical trials on glioma: a clinician's perspective. *Acta Neuropathol*. 2010;120(3):297–304.
13. Presant CA, Russell W, Alexander R, Fu Y. Soft-tissue and bone sarcoma histopathology peer review: the frequency of disagreement in diagnosis and the need for second pathology opinions. The Southeastern Cancer Study Group experience. *J Clin Oncol*. 1986;4(11):1658–61.
14. Coco DP, Goldblum JR, Hornick JL, Lauwers GY, Montgomery E, Srivastava A, et al. Interobserver variability in the diagnosis of crypt dysplasia in Barrett esophagus. *Am J Surg Pathol*. 2011;35(1):45–54.
15. Feagan BG, Sandborn WJ, D'Haens G, Pola S, McDonald JW, Rutgeerts P, et al. The role of centralized reading of endoscopy in a randomized controlled trial of mesalamine for ulcerative colitis. *Gastroenterology*. 2013;145(1):149–57.
16. Mahaffey KW, Harrington RA, Akkerhuis M, Kleiman NS, Berdan LG, Crenshaw BS, et al. Disagreements between central clinical events committee and site investigator assessments of myocardial infarction endpoints in an international clinical trial: review of the PURSUIT study. *Trials*. 2001;2(4):187.
17. Klompas M. Interobserver variability in ventilator-associated pneumonia surveillance. *Am J Infect Control*. 2010;38(3):237–9.
18. O'Donnell CP, Kamlin COF, Davis PG, Carlin JB, Morley CJ. Interobserver variability of the 5-minute Apgar score. *J Pediatr*. 2006;149(4):486–9.
19. Mitra D, Connolly D, Jenkins S, English P, Birchall D, Mandel C, et al. Comparison of image quality, diagnostic confidence and interobserver variability in contrast enhanced MR angiography and 2D time of flight angiography in evaluation of carotid stenosis. *Br J Radiol*. 2006;79(939):201–7.
20. Rubenfeld GD, Caldwell E, Granton J, Hudson LD, Matthay MA. Interobserver variability in applying a radiographic definition for ARDS. *Chest*. 1999;116(5):1347–53.
21. Thiesse P, Ollivier L, Di Stefano-Louineau D, Négrier S, Savary J, Pignard K, et al. Response rate accuracy in oncology trials: reasons for interobserver variability. Groupe Français d'Immunotherapie of the Fédération Nationale des Centres de Lutte Contre le Cancer. *J Clin Oncol*. 1997;15(12):3507–14.
22. Gwyther S, Bolis G, Gore M, WtB Huinink, Verweij J, Hudson I, et al. Experience with independent radiological review during a topotecan trial in ovarian cancer. *Ann Oncol*. 1997;8(5):463–8.
23. Scott CB, Nelson JS, Farnan NC, Curran WJ Jr, Murray KJ, Fischbach AJ, et al. Central pathology review in clinical trials for patients with malignant glioma. A report of radiation therapy oncology group 83-02. *Cancer*. 1995;76(2):307–13.
24. Hopper KD, Kasales CJ, Van Slyke MA, Schwartz TA, TenHave TR, Jozefiak JA. Analysis of interobserver and intraobserver variability in CT tumor measurements. *AJR Am J Roentgenol*. 1996;167(4):851–4.
25. Bauknecht H-C, Romano VC, Rogalla P, Klingebiel R, Wolf C, Bornemann L, et al. Intra-and interobserver variability of linear and volumetric measurements of brain metastases using contrast-enhanced magnetic resonance imaging. *Investig Radiol*. 2010;45(1):49–56.
26. Hricak H, Gatsonis C, Coakley FV, Snyder B, Reinhold C, Schwartz LH, et al. Early invasive cervical cancer: CT and MR imaging in preoperative evaluation—ACRIN/GOG comparative study of diagnostic performance and interobserver variability. *Radiology*. 2007;245(2):491–8.
27. McErlean A, Panicek DM, Zabor EC, Moskowitz CS, Bitar R, Motzer RJ, et al. Intra-and interobserver variability in CT measurements in oncology. *Radiology*. 2013;269(2):451–9.
28. Wormanns D, Diederich S, Lentschig M, Winter F, Heindel W. Spiral CT of pulmonary nodules: interobserver variation in assessment of lesion size. *Eur Radiol*. 2000;10(5):710–3.
29. Aldape K, Simmons ML, Davis RL, Miike R, Wiencke J, Barger G, et al. Discrepancies in diagnoses of neuroepithelial neoplasms: the San Francisco bay area adult glioma study. *Cancer*. 2000;88(10):2342–9.
30. Pandolfino JE, Vakil NB, Kahrilas PJ. Comparison of inter- and intraobserver consistency for grading of esophagitis by expert and trainee endoscopists. *Gastrointest Endosc*. 2002;56(5):639–43.
31. Ibrahim GM, Weidauer S, Macdonald RL. Interobserver variability in the interpretation of computed tomography following aneurysmal subarachnoid hemorrhage. *J Neurosurg*. 2011;115(6):1191–6.
32. Gietema HA, Wang Y, Xu D, van Klaveren RJ, de Koning H, Scholten E, et al. Pulmonary nodules detected at lung cancer screening: interobserver variability of semiautomated volume measurements. *Radiology*. 2006;241(1):251–7.
33. Hersh CP, Washko GR, Jacobson FL, Gill R, Estepar RSJ, Reilly JJ, et al. Interobserver variability in the determination of upper lobe-predominant emphysema. *Chest*. 2007;131(2):424–31.
34. Scholten AN, Smit VT, Beerman H, van Putten WL, Creutzberg CL. Prognostic significance and interobserver variability of histologic grading systems for endometrial carcinoma. *Cancer*. 2004;100(4):764–72.
35. Suzuki C, Torkzad MR, Jacobsson H, Åström G, Sundin A, Hatschek T, et al. Interobserver and intraobserver variability in the response evaluation of cancer therapy according to RECIST and WHO-criteria. *Acta Oncol*. 2010;49(4):509–14.
36. Gregory RL. The intelligent eye. 1970.
37. Gregory RL. *Eye and Brain: The Psychology of Seeing*. 2nd ed. New York: McGraw-Hill; 1973.
38. Rock I. *The Logic of Perception*. Cambridge: MIT Press; 1983.
39. Kundel HL. History of research in medical image perception. *J Am Coll Radiol*. 2006;3(6):402–8.
40. Kundel HL, Nodine CF. Interpreting chest radiographs without visual search. *Radiology*. 1975;116(3):527–32.
41. Kundel HL, Nodine CF. A visual concept shapes image perception. *Radiology*. 1983;146(2):363–8.
42. Nodine CF, Kundel HL. Using eye movements to study visual search and to improve tumor detection. *RadioGraphics*. 1987;7(6):1241–50.
43. Manning D. *The Handbook of Medical Image Perception and Techniques*. 2010.
44. Manning DJ, Gale A, Krupinski EA. Perception research in medical imaging. *Br J Radiol*. 2005;78(932):683–5.
45. Garland LH. Studies on accuracy of diagnostic procedures. *AJR*. 1959;82:25–38.

46. Ford R, O'Neal M, Moskowitz S, Fraunberger JJJCT. Adjudication rates between readers in blinded independent central review of oncology studies. *J Clin Trials*. 2016;6:289.
47. Maskell G. Error in radiology-where are we now? *Br J Radiol*. 2019;92(1095):20180845.
48. Vos M, Uitdehaag B, Barkhof F, Heimans J, Baayen H, Boogerd W, et al. Interobserver variability in the radiological assessment of response to chemotherapy in glioma. *Neurology*. 2003;60(5):826–30.
49. Lee HJ, Kim EK, Kim MJ, Youk JH, Lee JY, Kang DR, et al. Observer variability of Breast Imaging Reporting and Data System (BI-RADS) for breast ultrasound. *Eur J Radiol*. 2008;65(2):293–8.
50. Guillery RW, Sherman SM. Thalamic relay functions and their role in corticocortical communication: generalizations from the visual system. *Neuron*. 2002;33(2):163–75.
51. Hermans R, Feron M, Bellon E, Dupont P, Van den Bogaert W, Baert AL. Laryngeal tumor volume measurements determined with CT: a study on intra-and interobserver variability. *Int J Radiat Oncol Biol Phys*. 1998;40(3):553–7.
52. Berbaum KS, Franken EA, Dorfman DD, Miller EM, Caldwell RT, Kuehn DM, et al. Role of faulty visual search in the satisfaction of search effect in chest radiography. *Acad Radiol*. 1998;5(1):9–19.
53. Berbaum KS, Franken EA, Dorfman DD, Miller EM, Krupinski EA, Kreinbring K, et al. Cause of satisfaction of search effects in contrast studies of the abdomen. *Acad Radiol*. 1996;3(10):815–26.
54. Busby LP, Courtier JL, Glastonbury CM. Bias in radiology: the how and why of misses and misinterpretations. *RadioGraphics*. 2018;38(1):236–47.
55. Gilbert CD, Li W. Top-down influences on visual processing. *Nat Rev Neurosci*. 2013;14(5):350–63.
56. Jung R. *Visual Perception and Neurophysiology. Central Processing of Visual Information A: Integrative Functions and Comparative Data*. Berlin: Springer; 1973. p. 296–301.
57. Borradaile K, Ford R, O'Nea M, Byrne K. Discordance between BICR readers. *Appl Clin Trials*. 2010;19(11).
58. Kim YW, Mansfield LT. Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors. *Am J Roentgenol*. 2014;202(3):465–70.
59. MGH. T*he Clinical Trials Program in the Mass General Department of Radiology Provides Access to the Expertise and Technology of a Premiere Academic Radiology Department*. 2020 https://www.massgeneral.org/imaging/approach/professional-services/radiology-clinical-trials.
60. Scher HI, Morris MJ, Stadler WM, Higano CS, Halabi S, Smith MR, et al. The Prostate Cancer Working Group 3 (PCWG3) consensus for trials in castration-resistant prostate cancer (CRPC). *Am Soc Clin Oncol*. 2015. https://doi.org/10.1200/jco.2015.33.15_suppl.5000.
61. Scher HI, Halabi S, Tannock I, Morris M, Sternberg CN, Carducci MA, et al. Design and end points of clinical trials for patients with progressive prostate cancer and castrate levels of testosterone: recommendations of the Prostate Cancer Clinical Trials Working Group. *J Clin Oncol*. 2008;26(7):1148.
62. Lee HJ, Goo JM, Lee CH, Park CM, Kim KG, Park E-A, et al. Predictive CT findings of malignancy in ground-glass nodules on thin-section chest CT: the effects on radiologist performance. *Eur Radiol*. 2009;19(3):552–60.
63. Miglioretti DL, Gard CC, Carney PA, Onega TL, Buist DS, Sickles EA, et al. When radiologists perform best: the learning curve in screening mammogram interpretation. *Radiology*. 2009;253(3):632–40.
64. Tucker L, Gilbert FJ, Astley SM, Dibden A, Seth A, Morel J, et al. Does reader performance with digital breast tomosynthesis vary according to experience with two-dimensional mammography? *Radiology*. 2017;283(2):371–80.
65. Wassberg C, Akin O, Vargas HA, Shukla-Dave A, Zhang J, Hricak H. The incremental value of contrast-enhanced MRI in the detection of biopsy-proven local recurrence of prostate cancer after radical prostatectomy: effect of reader experience. *Am J Roentgenol*. 2012;199(2):360–6.
66. Sridhara R, Mandrekar SJ, Dodd LE. Missing data and measurement variability in assessing progression-free survival endpoint in randomized clinical trials. *AACR*; 2013.
67. Dodd LE, Korn EL, Freidlin B, Jaffe CC, Rubinstein LV, Dancey J, 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(22):3791.
68. Bogaerts J, Ford R, Sargent D, Schwartz LH, Rubinstein L, Lacombe D, et al. Individual patient data analysis to assess modifications to the RECIST criteria. *Eur J Cancer*. 2009;45(2):248–60.
69. Muenzel D, Engels H-P, Bruegel M, Kehl V, Rummeny EJ, Metz S. Intra- and inter-observer variability in measurement of target lesions: implication on response evaluation according to RECIST 1.1. *Radiol Oncol*. 2012;46(1):8–18.
70. Ishimori T, Patel PV, Wahl RL. Detection of unexpected additional primary malignancies with PET/CT. *J Nucl Med*. 2005;46(5):752–7.
71. Wiggermann V, Hernandez-Torres E, Traboulsee A, Li D, Rauscher A. FLAIR2: a combination of FLAIR and T2 for improved MS lesion detection. *Am J Neuroradiol*. 2016;37(2):259–65.
72. Eisenhauer EA, Therasse P, Bogaerts J, Schwartz LH, Sargent D, Ford R, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45(2):228–47.
73. Moertel CG, Hanley JA. The effect of measuring error on the results of therapeutic trials in advanced cancer. *Cancer*. 1976;38(1):388–94.
74. Barrington SF, Mikhaeel NG, Kostakoglu L, Meignan M, Hutchings M, Müeller SP, et al. Role of imaging in the staging and response assessment of lymphoma: consensus of the international conference on malignant lymphomas imaging working group. *J Clin Oncol*. 2014;32(27):3048–58.
75. Hasenclever D, Kurch L, Mauz-Körholz C, Elsner A, Georgi T, Wallace H, et al. qPET—a quantitative extension of the Deauville scale to assess response in interim FDG-PET scans in lymphoma. *Eur J Nuclear Med Mol Imaging*. 2014;41(7):1301–8.
76. Meignan M, Itti E, Gallamini A, Younes A. FDG PET/CT imaging as a biomarker in lymphoma. *Eur J Nuclear Med Mol Imaging*. 2015;42(4):623–33.
77. Nols N, Mounier N, Bouazza S, Lhommel R, Costantini S, Vander Borght T, et al. Quantitative and qualitative analysis of metabolic response at interim positron emission tomography scan combined with International Prognostic Index is highly predictive of outcome in diffuse large B-cell lymphoma. *Leukemia Lymphoma*. 2014;55(4):773–80.
78. Beaumont H, Evans TL, Klifa C, Guermazi A, Hong SR, Chadjaa M, et al. Discrepancies of assessments in a RECIST 1.1 phase II clinical trial–association between adjudication rate and variability in images and tumors selection. *Cancer Imaging*. 2018;18(1):50.
79. Oxnard GR, Zhao B, Sima CS, Ginsberg MS, James LP, Lefkowitz RA, et al. Variability of lung tumor measurements on repeat computed tomography scans taken within 15 minutes. *J Clin Oncol*. 2011;29(23):3114–9.
80. Li Q, Gavrielides MA, Sahiner B, Myers KJ, Zeng R, Petrick N. Statistical analysis of lung nodule volume measurements with CT in a large-scale phantom study. *Med Phys*. 2015;42(7):3932–47.

81. Erasmus JJ, Gladish GW, Broemeling L, Sabloff BS, Truong MT, Herbst RS, et al. Interobserver and intraobserver variability in measurement of non–small-cell carcinoma lung lesions: implications for assessment of tumor response. *J Clin Oncol*. 2003;21(13):2574–82.
82. Cornelis FH, Martin M, Saut O, Buy X, Kind M, Palussiere J, et al. Precision of manual two-dimensional segmentations of lung and liver metastases and its impact on tumour response assessment using RECIST 1.1. *Eur Radiol Exp*. 2017;1(1):16.
83. Alpert HR, Hillman BJ. Quality and variability in diagnostic radiology. *J Am Coll Radiol*. 2004;1(2):127–32.
84. Robert NJ, Diéras V, Glaspy J, Brufsky AM, Bondarenko I, Lipatov ON, et al. RIBBON-1: randomized, double-blind, placebo-controlled, phase III trial of chemotherapy with or without bevacizumab for first-line treatment of human epidermal growth factor receptor 2–negative, locally recurrent or metastatic breast cancer. *J Clin Oncol*. 2011;29(10):1252–60.
85. FDA. *FDA Briefing Document Oncologic Drugs Advisory Committee Meeting-ucm250378*. UDoHaH, editor. Rockville, MD2018. April 12, 2011.
86. Raunig D, Goldmacher G, Conklin J. *Local Evaluation and Blinded Central Review Comparison: A Victim of Meta-analysis Shortcomings*. Los Angeles: SAGE Publications Sage CA; 2013.
87. Zhang JJ, Chen H, He K, Tang S, Justice R, Keegan P, et al. Evaluation of blinded independent central review of tumor progression in oncology clinical trials: a meta-analysis. *Ther Innov Regul Sci*. 2013;47(2):167–74.
88. Martí-Bonmatí L, Alberich-Bayarri Á, Ladenstein R, Blanquer I, Segrelles JD, Cerdá-Alberich L, et al. PRIMAGE project: predictive in silico multiscale analytics to support childhood cancer personalised evaluation empowered by imaging biomarkers. *Eur Radiol Exp*. 2020;4:1–11.
89. Berbaum KS, Franken EA, Honda H, McGuire C, Weis RR, Barloon T. Evaluation of a PACS workstation for assessment of body CT studies. *J Comput Assist Tomogr*. 1990;14(5):853–8.
90. Beam CA, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists: findings from a national sample. *JAMA Internal Med*. 1996;156(2):209–13.
91. Krupinski EA. The future of image perception in radiology: synergy between humans and computers. *Acad Radiol*. 2003;10(1):1–3.
92. Degnan AJ, Ghobadi EH, Hardy P, Krupinski E, Scali EP, Stratchko L, et al. Perceptual and interpretive error in diagnostic radiology—causes and potential solutions. *Acad Radiol*. 2019;26(6):833–45.
93. FDA. *Conduct of Clinical Trials of Medical Products During the COVID-19 Public Health Emergency, Guidance for Industry Investigators, and Institutional Review Boards*. In: Services UDoHaH, editor. Rockville, MD2020.