

RADPEER™ Scoring White Paper

Valerie P. Jackson, MD^a, Trudie Cushing, MS^b, Hani H. Abujudeh, MD, MBA^c,
James P. Borgstede, MD^d, Kenneth W. Chin, MD^e, Charles K. Grimes, MD^f,
David B. Larson, MD^g, Paul A. Larson, MD^h, Robert S. Pyatt Jr, MDⁱ,
William T. Thorwarth Jr, MD^j

The ACR's RADPEER™ program began in 2002; the electronic version, e-RADPEER™, was offered in 2005. To date, more than 10,000 radiologists and more than 800 groups are participating in the program. Since the inception of RADPEER, there have been continuing discussions regarding a number of issues, including the scoring system, the subspecialty-specific subcategorization of data collected for each imaging modality, and the validation of interfacility scoring consistency. This white paper reviews the task force discussions, the literature review, and the new recommended scoring process and lexicon for RADPEER.

Key Words: Peer review, medical errors, harm score, undercalls, overcalls, misinterpretations, disagreement rates

J Am Coll Radiol 2009;6:21-25. Copyright © 2009 American College of Radiology

INTRODUCTION

The ACR established a task force on patient safety in 2000, in response to the 1999 Institute of Medicine report *To Err Is Human* [1], which estimated that as many as 98,000 people die in hospitals as a result of preventable medical errors. Although medical imaging was not cited as an area of practice with high error rates, the ACR's task force established several committees to address patient safety issues. One of the committees in this task force addressed model peer review and self-evaluation. That committee developed the RADPEER™ program, a radiology peer-review process, and conducted a pilot of the program at 14 sites in 2001 and 2002. After the pilot study, the program was offered to ACR members in 2002.

RADPEER was designed to be a simple, cost-effective process that allows peer review to be performed during

the routine interpretation of current images. If prior images and reports are available at the time a new study is being interpreted, these prior studies and the accuracy of their interpretation would typically be evaluated at the time the radiologist interprets the current study. In addition, at the time of the interpretation of the current study, the radiologist may have additional information that is helpful in assessing the interpretation of the prior study. This may include the progression or regression of findings on the current imaging study or additional history, including the findings of intervening nonimaging studies or procedures. The process requires no additional interpretive work on the part of radiologists beyond what already is currently being done. RADPEER simply creates a system that allows "old images" and "old interpretations" to be collected and structured in a reviewable format. The accuracy of prior reports is scored by the current interpreter of the new study using a standardized, 4-point rating scale (Table 1).

Although this scoring system has worked well for the past 5 years, there has been continued confusion over the meaning of some categories. Scores 1 and 4 are easy to understand. However, score 3 does not mention misinterpretation or disagreement and could potentially be used in a situation in which an image was correctly interpreted but the reviewer merely felt that it was an easy diagnosis. Likely the most confusing is score 2, "difficult diagnosis, not ordinarily expected to be made." It is unclear whether there is an actual disagreement with the original interpretation or if the score is being used because it was a great pickup. Scores of 1 and 2 require no action, but scores of 3 and 4 require internal review by

^aIndiana University School of Medicine, Indianapolis, Indiana.

^bAmerican College of Radiology, Reston, Virginia.

^cMassachusetts General Hospital, Boston, Massachusetts.

^dUniversity of Colorado, Denver, Colorado.

^eSan Fernando Valley Interventional Radiology & Imaging Center, Encino, California.

^fMaine Medical Center, Portland, Maine.

^gThe Children's Hospital, Aurora, Colorado.

^hRadiology Associates of Fox Valley, Neenah, Wisconsin.

ⁱChambersburg Imaging Associates, Chambersburg, Pennsylvania.

^jCatawba Radiological Associates, Hickory, North Carolina.

Corresponding author and reprints: Valerie P. Jackson, MD, Indiana University School of Medicine, Department of Radiology, Room 0663, 550 University Blvd, Indianapolis, IN 46202-5149; e-mail: vjackson@iupui.edu.

Table 1. Current RADPEER scoring system

Score	Meaning
1	Concur with interpretation
2	Difficult diagnosis, not ordinarily expected to be made
3	Diagnosis should be made most of the time
4	Diagnosis should be made almost every time—misinterpretation of findings

the local peer-review committee to validate or change, if necessary, the original RADPEER score.

Each institution (radiology group) is assigned a unique identifier number. To maintain confidentiality, facilities assign each physician a numeric identifier (such as 101) to use when information is submitted to the ACR. The actual names of the participating radiologists are not provided to the ACR.

RADPEER scoring was originally performed using machine readable cards; in 2005, a Web-based program, e-RADPEER™, was established. Completed cards or electronic scores are submitted to the ACR, and reports are generated that provide

- summary statistics and comparisons for each radiologist by modality,
- summary data for each facility by modality, and
- data summed across all participating facilities.

The reports should demonstrate trends that radiologists may use to focus their continuing medical education activities. Efforts to optimize interpretive skills should result in improvements in patient care.

The original model peer review committee members reviewed several documents, including examples of scoring from the literature [2,3] (W. Thorwarth, personal communication) and samples submitted from committee members' own practices. Everyone agreed that for any program to be effective and widely accepted, it needs to be simple and user friendly. The committee members reviewed several examples of scoring and eventually decided on the 4-point system shown in Table 1. Although there was discussion regarding the inclusion of "clinically significant" in the scoring language, the committee steered away from this because of the difficulty in tracking a case for evidence of clinical significance and outcome. The committee also discussed the categories selected and whether modality or body system was preferable. Because the original card-based system used in RADPEER required radiologists to manually enter data onto cards, the need to keep the categories and scoring simple was emphasized.

RADPEER participation has increased every year, with substantial growth in 2007 after the ACR's mandate that all sites applying for any of the voluntary accreditation programs (computed tomography, magnetic resonance, ultrasound, positron emission tomography, nuclear medicine, and breast ultrasound) have evidence of a physician peer-review program, either RADPEER or their own internal programs. The number of participating radiologists has grown to more than 10,000.

The summarized RADPEER data collected through December 2007 are shown in Table 2. These data raise important questions. Fewer than 0.5% of the scores are 3 or 4. Does this reflect the quality of the interpretive skills of radiologists or a reluctance to assign less than perfect scores to colleagues? However, the RADPEER data are similar to those reported in the literature. For example, Soffa et al [4] found a disagreement rate of 3.48%. Combining RADPEER scores of 2, 3, and 4 gives a total disagreement rate of 2.91%. If the scores are not a true reflection of individual radiologists' interpretive skills, does the RADPEER process serve as a tool for improving patient safety or continuous quality improvement?

Since the inception of RADPEER, there have been continuing discussions regarding a number of issues, including the scoring system, the subspecialty-specific subcategorization of data collected for each imaging modality, and the validation of interfacility scoring consistency. In addition, there has been controversy regarding the inclusion of the clinical significance for scores of 2, 3, and 4. When RADPEER was originally developed, the committee members felt that adding clinical significance would require follow-up that either could not be done or would place an additional burden on radiology resources.

TASK FORCE ON RADPEER SCORING

Because of the issues regarding the RADPEER scoring process, a task force was formed to review the literature and various scoring methods to determine if a change would be warranted. The task force met on September 15, 2007, and consisted of members of the RADPEER committee, representatives from ACR leadership, and a radiology resident. The task force members reviewed the

Table 2. Summary of RADPEER scores

Score	Percentage of Total Scored Cases
1	97.11
2	2.51
3	0.32
4	0.07

Table 3. Melvin et al [6] scoring system of discrepancy

Grade	Significance
0 = No discrepancy	0 = None
1 = Minor	1 = Minor (incidental to treatment/management)
2 = Significant	2 = Significant (affects treatment/management, not outcome)
3 = Major	3 = Major (affects outcome)

current language, the literature, and several proposed changes to the current scoring system.

There were several suggestions that the scoring system be changed to improve the ease of scoring, improve the consistency of scoring, and reflect the clinical significance of the various levels of disagreement. This would move RADPEER from a scoring system based on standards of care to an outcomes-based system, more in line with peer-reviewed systems described in the literature [4-6]. In the current climate of emphasis on patient safety, this could perhaps become a more durable product.

The task force reviewed the literature on peer-review processes in medicine and specifically in radiology. Lee et al [5] described a 5-point system, with 1 representing “not significant” and 5, “highly significant,” causing delays in diagnosis (false-negative results), unwarranted invasive procedures (false-positive results), or incorrect treatments. Similarly, Melvin et al [6] also used a grading system that combined the severity of a discrepancy with its clinical significance (Table 3).

Clinical Significance

If RADPEER is to have an impact on patient safety, should the clinical significance of a score of 2, 3, or 4 be evaluated, as has been done in other peer-review processes? The task force members agreed that it would be useful to provide the reviewer an option to not only assign a score but also evaluate clinical significance. Thus,

in the proposed system, for categories 2 to 4, the reviewer has the option to check the items “unlikely to be clinically significant” and “likely to be clinically significant.”

There was concern expressed about the ability to rank something as “clinically significant,” because that cannot be ascertained in all cases. One of the task force members reported that his group originally had a “clinically significant” category but discontinued its use because of difficulty with scoring. For example, if a radiologist misses a metastatic lesion in a patient with other metastases, it would not be as clinically significant as a lesion in a patient with no known disease. Many task force members felt that the determination of clinical significance does not have to be a difficult process based on absolute outcome measures. Instead, a “gut” assessment of the likelihood of impact of the discrepancy on patient care would be adequate.

In any acceptable peer-review program under the ACR’s accreditation requirements (Appendix A), there must be “reviewer assessment of the agreement of the original report with a subsequent review” and “policies and procedures for action to be taken on significantly discrepant peer review findings for purposes of achieving quality outcomes improvements.” The whole point of peer review is to compare studies to assess reviewer accuracy and, should discrepancies exist, having a system in place to assess the need for reviewer improvement that should ultimately improve patient care.

After discussion of the issues regarding the meaning of each score and clinical significance, the task force members agreed on a new RADPEER scoring language (Table 4). The scoring numbers remain the same, but some of the definitions have changed. The committee members felt strongly that examples of cases for each of the scores are necessary to clarify the scoring process for radiologists (Appendix B). Score 2 was better defined to indicate that it represents a discrepancy in interpretation, but for a finding difficult enough that it is an understandable miss. The task force discussed changing score 3 to “substantial discrepancy in interpretation” and score 4 to “major discrepancy in interpretation.” However, the ACR’s legal

Table 4. Proposed RADPEER scoring language

Score	Meaning	Optional
1	Concur with interpretation	
2	Discrepancy in interpretation/not ordinarily expected to be made (understandable miss)	a. Unlikely to be clinically significant b. Likely to be clinically significant
3	Discrepancy in interpretation/should be made most of the time	a. Unlikely to be clinically significant b. Likely to be clinically significant
4	Discrepancy in interpretation/should be made almost every time—misinterpretation of finding	a. Unlikely to be clinically significant b. Likely to be clinically significant

staffers felt that the terms *substantial* and *major* are too vague and recommended that the original language be maintained, changing the word *diagnosis* to *interpretation*.

The current system language is more related to misses or standard of care, whereas the proposed system is more widely applicable and outcomes based, similar to the “harm score” of the Pennsylvania Patient Safety Reporting System [7]. The harm score (with 10 categories) ranges from circumstances that could cause adverse events to events that contributed to or resulted in death. The Joint Commission also looks at “harm vs no detectable harm” [8], assessing the impact of medical errors and systems failure, commonly referred to as harm to patients. In addition, the proposed scoring system addresses the issues of both undercalls (misses) and overcalls (which may lead to unnecessary additional tests or intervention).

Some members expressed concern that a major change to the language would cause all of the accumulated data to be lost. The ACR’s Research Department reviewed the proposed scoring system and felt that any change should preserve comparability with respect to the distinction of a score of 1 vs any other score. The proposed system does this in large part by maintaining a 4-category scoring language. Because the concern is that categories 2 to 4 have been used inconsistently, the change should preserve what is relatively reliable in the historical data. They suggested that the task force may want to recommend that the terminology be reevaluated and change considered about every 5 years.

Legal Implications of RADPEER Language

Obviously, the improvement of patient care includes taking action when it is discovered that a study was misread, thereby affecting appropriate patient care. Failure to take the appropriate action can subject health care providers to malpractice liability as easily as the initial misread of the study. Moreover, failure to act on a misread can be viewed by a jury as reason to impose punitive sanctions on a physician, resulting in a malpractice conviction and a higher judgment than might have been awarded for the failure to make the initial interpretation. Thus, regardless of the language used in the RADPEER definitions, the liability exposure remains the same. The current RADPEER system requires that all scores of 3 and 4 be reviewed by the local peer-review committee (the group’s internal peer-review process or committee) for validation and appropriate action.

Validation of Scores

The validation of RADPEER, through the development of some type of process to standardize scoring so that scoring at one facility is comparable with scoring at another, was discussed at length. There are several concerns with a validation process. It is likely that any validation

process would require tracking radiologists’ identities and the possible loss of anonymity and perhaps protection from discovery. If the ACR develops a model whereby an outside “expert” would overread the score of the original RADPEER reviewer, how would it be determined who is expert enough and adequately trained to determine the correct score? In addition, any type of validation process would involve additional cost and resources for both the ACR and the facility. Thus, the RADPEER committee will study the validation issue. It was felt to be out of the scope of the task force at this time.

CONCLUSION

In summary, the task force is proposing a scoring system that will build on the current system, maintaining the current system of numbers for scoring but making the categories clearer. In addition, radiologists would have the option to give their opinions regarding the clinical significance of discrepancies in interpretation, more in keeping with other peer-review methods described in the literature. The task force members all strongly agreed that better explanation of the scoring, with examples, is necessary to help standardize the scoring method among participants. In addition, any future changes should be accompanied by changes to the lexicon. Future groups or task forces that discuss RADPEER language and scoring should be mindful of the impact of any changes on existing statistical data. It is probably wise to expand the existing scoring system rather than change to a completely new system, so as not to lose previous data.

REFERENCES

1. Kohn LT, Corrigan JM, Donaldson MS, editors. To err is human: building a safer health system. Washington, DC: National Academy Press; 2000.
2. Committee on Quality Assurance, Commission on Standards and Accreditation. A guide to continuous quality improvement in medical imaging, Reston, Va: American College of Radiology; 1996.
3. AuntMinnie.com. Integrating peer review into workstation eases productivity concerns. Available at: <http://www.auntminnie.com/index.asp?sec=news&sub=mtf&itemid=956>. Accessed June 20, 2000.
4. Soffa J, Lewis RS, Sunshine JH, Bhargavan M. Disagreement in interpretation: a method for the development of benchmarks for quality assurance in imaging. *J Am Coll Radiol* 2004;1:212-7.
5. Lee KT. Quality—a radiology imperative: interpretation accuracy and pertinence. *J Am Coll Radiol* 2007;4:162-5.
6. Melvin C, Bodley R, Booth A, Meagher T, Record C, Savage P. Managing errors in radiology: a working model. *Clin Radiol* 2004;59:841-5.
7. Pennsylvania Patient Safety Authority. Pennsylvania Patient Safety Reporting System training manual and users guide. Harrisburg, Pa: Pennsylvania Patient Safety Authority; 2004:47.
8. Chang A, Schyve PM, Croteau RJ, O’Leary DS, Loeb JM. The JCAHO patient safety event taxonomy: a standardized terminology and classification schema for near misses and adverse events. *Int J Quality Health Care*. Available at: <http://intqhc.oxfordjournals.org/cgi/reprint/mzi021v1.pdf>. Accessed November 20, 2008.

APPENDIX A

Items That Must Be Included in an Acceptable Alternative Physician Peer-Review Program

- A peer-review process that includes a double reading with 2 physicians interpreting the same study
- A peer-review process that allows for the random selection of studies to be reviewed on a regularly scheduled basis
- Examinations and procedures representative of the work of each physician's specialty
- Reviewer assessment of the agreement of the original report with subsequent review (or with surgical or pathologic findings)
- A classification of peer-review findings with regard to level of quality concerns (eg, a 4-point scoring scale)
- Policies and procedures for action to be taken on significantly discrepant peer-review findings for the purpose of achieving quality outcomes improvement
- Summary statistics and comparisons generated for each physician by modality
- Summary data for each facility or practice by modality

APPENDIX B

Examples of Scoring

Note: Scoring should include both primary findings and incidental findings on the imaging study. Both misses and overcalls can be included.

Score of 1: "Concur with original reading": self-explanatory

Score of 2: "Discrepancy in interpretation/not ordinarily expected to be made (understandable miss)"

A. "Unlikely to be clinically significant"

- Small knee collateral ligament tear (ie, subtle or difficult to appreciate finding)
- Osteopoikilosis that is not clinically significant (ie, esoteric finding)
- 7-mm mesenteric lymph node on abdominal computed tomography (CT)
- Small (5-mm) apical pneumothorax on overpenetrated portable chest radiography after subclavian line placement
- Minimally calcified (<3 cm) abdominal aortic aneurysm on kidney, ureter, and bladder scan
- Old, healed long-bone fracture (ie, apparent on single view)
- Subtle mass (probable benign lymph node) on mammography

B. "Likely to be clinically significant"

- Subtle or early lung cancer seen on chest CT in retrospect (ie, difficult to diagnose prospectively)

- Subtle meningeal enhancement on brain CT or magnetic resonance imaging (MRI)
- Small subdural hematoma around cerebellar tentorium
- Subtle scapholunate separation
- Small minimally radiopaque soft-tissue glass foreign body on hand radiography
- Subtle 1.5-cm pancreatic tail mass
- Early vascular calcifications on screening mammography, recalled for additional imaging (overcall)

Score of 3: "Discrepancy in interpretation/should be made most of the time"

A. "Unlikely to be clinically significant"

- 2-cm bone cyst noted on knee MRI
- Pneumoperitoneum on abdominal film of patient one day after abdominal surgery
- Vertebral body hemangioma on spine MRI
- 3-cm thyroid mass on chest CT
- 5-mm calcified renal calculus without associated hydronephrosis on computed tomographic urography

B. "Likely to be clinically significant"

- Small subdural hematoma on brain CT
- Skin fold interpreted as pneumothorax in newborn with subsequent placement of chest tube
- Asymmetric 2-cm breast mass on chest CT
- 2-cm para-aortic or pelvic lymph node
- Periappendiceal or pericolic fat stranding
- 1.5-cm adrenal mass in patient with lung mass
- Cluster of pleomorphic microcalcifications on mammography
- Pericardial effusion on chest CT
- Short single-segment Crohn's disease on small bowel follow-through examination
- Lateral meniscus tear on knee MRI

Score of 4: "Discrepancy in interpretation/should be made almost every time—misinterpretation of finding"

A. "Unlikely to be clinically significant"

- 4-cm pelvic lymph node in patient beginning chemotherapy for lymphoma
- 2-cm calcified gallstone on CT of a patient with lower left quadrant pain and diverticulitis

B. "Likely to be clinically significant"

- Displaced fracture of base of fifth metatarsal
- 25% slipped capital femoral epiphysis in 12-year-old patient
- Tension pneumothorax
- Large medial meniscus tear on knee MRI
- 3-cm hilar lymph node on chest CT
- 2-cm lung nodule on chest radiography
- "Classic" molar pregnancy on pelvic ultrasound
- Obvious hamartoma on mammography for which biopsy was recommended (overcall)