BJR

Received: 30 April 2013 | Revised: 28 June 2013 | Accepted: 2 July 2013

doi: 10.1259/bjr.20130245

Cite this article as:
Mucci B, Murray H, Downie A, Osborne K. Interrater variation in scoring radiological discrepancies. Br J Radiol 2013;86:20130245.

FULL PAPER

Interrater variation in scoring radiological discrepancies

[1]B MUCCI, FRCR, [2]H MURRAY, MSc, [1]A DOWNIE, FRCR and [1]K OSBORNE, FRCR

[1]Department of Radiology, South Glasgow University Hospitals, Southern General Hospital, Glasgow, Scotland, UK
[2]Robertson Centre for Biostatistics, University of Glasgow, Glasgow, Scotland, UK

Address correspondence to: Dr Brian Mucci
E-mail: *Brian.Mucci@ggc.scot.nhs.uk*

Objective: Discrepancy meetings are an important aspect of clinical governance. The Royal College of Radiologists has published advice on how to conduct meetings, suggesting that discrepancies are scored using the scale: 0=no error, 1=minor error, 2=moderate error and 3=major error. We have noticed variation in scores attributed to individual cases by radiologists and have sought to quantify the variation in scoring at our meetings.

Methods: The scores from six discrepancy meetings totalling 161 scored events were collected. The reliability of scoring was measured using Fleiss' kappa, which calculates the degree of agreement in classification.

Results: The number of cases rated at the six meetings ranged from 18 to 31 (mean 27). The number of raters ranged from 11 to 16 (mean 14). Only cases where all the raters scored were included in the analysis. The Fleiss' kappa statistic ranged from 0.12 to 0.20, and mean kappa was 0.17 for the six meetings.

Conclusion: A kappa of 1.0 indicates perfect agreement above chance and 0.0 indicates agreement equal to chance. A rule of thumb is that a kappa ≥0.70 indicates adequate interrater agreement. Our mean result of 0.172 shows poor agreement between scorers. This could indicate a problem with the scoring system or may indicate a need for more formal training and agreement in how scores are applied.

Advances in knowledge: Scoring of radiology discrepancies is highly subjective and shows poor interrater agreement.

Discrepancy in radiological practice is inevitable. Although the technology of image acquisition has advanced rapidly in recent years, the final radiological opinion is still inevitably the product of individual radiologists. Human errors, particularly observation and interpretation errors, are unavoidable. A regular review of radiological discrepancies is now undertaken in most radiology departments. In 2007, the Royal College of Radiologists (RCR) issued advice [1] and stated that "some errors are greater than others". It was suggested that incidents be scored to indicate the "grade" or seriousness of the discrepancy. Scoring as in Table 1 was suggested.

This is similar to that proposed by Melvin et al [2]. Although a single score is mentioned in the RCR document, Melvin et al suggested giving separate scores to the degree of radiological discrepancy and to the significance of the discrepancy in terms of practical outcome. Similar scoring systems, including the American College of Radiologists web-based system RADPEER™ [3] and other semi-commercial systems such as peerView (peerView Inc., Sarastosa, FL), have been proposed elsewhere in the world. The literature is unclear as to how such gradings should be used [4–6]. There is potential for discrepancies graded as serious to lead to questions regarding a radiologist's competence and ability to practice. Discrepancy scoring is used by some commercial teleradiology companies [7], and it has been advocated by Hussain et al [8] that a certain level of discrepancy should lead to "restricted privileges or termination". In the United Kingdom, the General Medical Council (GMC) has stated that, as part of relicensing, doctors should "regularly participate in activities that review and evaluate the quality of your work" and that "activities should be robust, systematic and relevant to your work. They should include an element of evaluation and action" [9]. In this context, it is vital that any grading of radiological discrepancy should be robust and reproducible. In our department, we review discrepancies as recommended by the RCR, and in this study, we have attempted to measure the interrater reliability of a group of radiologists in applying a score to radiological discrepancy.

METHODS AND MATERIALS

South Glasgow Hospitals conduct radiological discrepancy meetings on a regular basis at two main sites (Southern General Hospital and Victoria Infirmary). Incidents are collected, prepared and presented by a discrepancy convenor as laid down by the RCR guidance [1]. All incidents are presented with patient and radiologist identification removed. The clinical information and the radiological report are provided. Each incident is scored by consultant radiologists attending the presentation using the four-point

Table 1. Scoring grades

Rating score	Meaning
0	No discrepancy
1	Minor discrepancy
2	Significant discrepancy
3	Major discrepancy

scale in Table 1. Instructions were that the score was for degree of discrepancy. Scoring was done anonymously using paper forms returned at the end of the meeting. Doctors in training were encouraged to return scores, but these are not used for individual feedback to reporting radiologists and were excluded from this study. We reviewed the scores from six consecutive meetings comprising 161 scored discrepancies.

Statistical methods

The %MAGREE macro for SAS v. 9.3 (SAS Institute, Cary, NC) was used to calculate interrater agreement between the multiple raters for each consultant radiological meeting. This macro is based on Fleiss' kappa statistic [10]. The strength of agreement for the kappa statistics was interpreted using the scale proposed by Landis and Koch [11]. A kappa value <0 would be considered to be no agreement, 0.01–0.20 slight agreement, 0.21–0.40 fair agreement, 0.41–0.60 moderate agreement, 0.61–0.80 substantial agreement and 0.81–1.00 almost perfect agreement.

RESULTS

The number of cases rated at the six meetings ranged from 18 to 31 (mean 27). The number of raters ranged from 11 to 16 (mean 14). Only cases where all the raters scored were included in the analysis. A total of 11 cases were excluded from the 6 meetings (7 cases from Meeting 2, 3 cases from Meeting 3 and 1 case from Meeting 4). The Fleiss' kappa statistic ranged from 0.12 to 0.20 (mean 0.17) for the 6 meetings (Table 1). All meetings were found to have only slight interrater agreement. For Meetings 2, 3, 4 and 6, the agreement was strongest for a score of 0 (*i.e.* no discrepancy) with the kappa statistics ranging from 0.25 to 0.42, and for Meetings 1 and 5, the agreement was strongest for a score of 1 (*i.e.* minor error) with kappa statistics of 0.23 and 0.19, respectively. For Meetings 1–5, no outlying raters were found with all raters giving a median score of 1 or 2. For Meeting 6, one rater had a median score of 3. Excluding this rater from analysis did not change the overall results; the kappa statistic was still found to have only slight interrater agreement [kappa statistic 0.19 (standard error 0.015)]. Results are summarised in Table 2.

Table 2. Interrater agreement

Meeting	Number of cases	Number of raters	Kappa (standard error)
Meeting 1	26	16	0.201 (0.0138)
Meeting 2	29	15	0.177 (0.0115)
Meeting 3	30	15	0.183 (0.0123)
Meeting 4	27	11	0.184 (0.0172)
Meeting 5	31	14	0.122 (0.0125)
Meeting 6	18	16	0.163 (0.0137)

We found peer review of radiological discrepancies to be subjective. As an illustration, we present three cases from our meetings:

Case 1: A middle-aged female underwent an ultrasound scan, which was reported as normal (Figure 1a). CT scan a few days later detected a large renal mass (Figure 1b). Scores ranged from 0 (no error) to 3 (serious error), with a median score of 2 (moderate error). The discrepancy is clear but the difficulty in assessing static ultrasound images was raised by many at the meeting.

Case 2: An elderly post-operative patient had a CT pulmonary angiogram reported as normal; the next day, the patient underwent contrast CT abdomen owing to continued pleuritic pain. Subphrenic collection in the left upper quadrant was visible on the initial study (Figure 2a), but was better seen on subsequent examination (Figure 2b). This was scored as 2 (moderate error) by most readers but as 3 (serious error) by 3 and as 1 (minor error) by 2. The median score was 2.

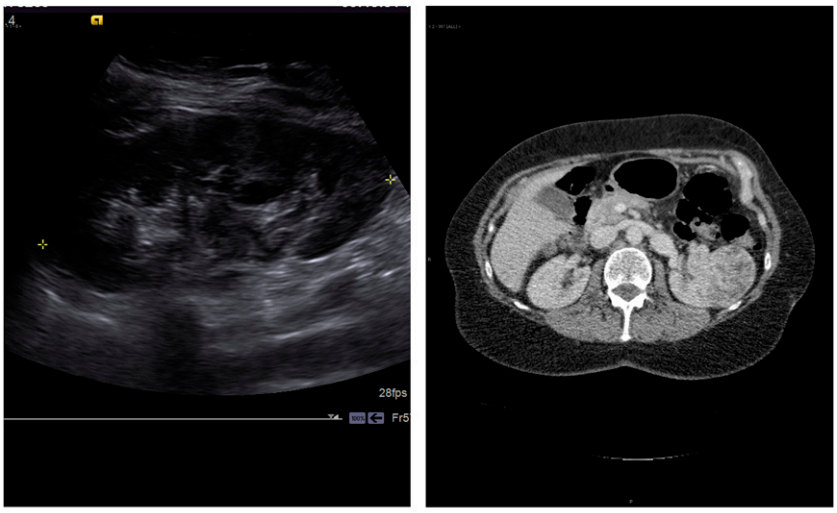
Case 3: An anteroposterior chest radiograph (Figure 3a) was reported as pulmonary shadowing, with no additional finding. CT scan within a few days (Figure 3b) revealed extensive lymphadenopathy. Scores varied from 1 (minor error) by 3 scorers, 2 (moderate error) by 6 scorers and 3 (serious error) by 1 scorer with an additional scorer giving 1.5! This demonstrates a wide range of opinions as to the severity of the error.

DISCUSSION

Discrepancy meetings as proposed by Melvin et al [2] and as recommended by the RCR [1] are an important and effective forum for review of errors and complications in radiology. The selection of errors is variable, and those scored depend on notification to the error coordinator, but our study reflects those notified in a large department. In the context of such a meeting, attempts to grade errors seem desirable. It is estimated that 4% of radiologists' daily work will contain errors [4]. Authors such as Berlin [12] have produced multiple papers describing incidents and their legal consequences. Although imaging technology has made considerable advances in the past few decades, the interpretation of images remains based on the observational and interpretational skills of human observers. The gap between image acquisition ability and our ability to interpret them has widened. Berlin [13] has pointed out how unsuccessful we have been at reducing error.

Medicine, in general, and radiology, in particular, have attempted to analyse errors in the same way that the aviation industry reviews accidents and near misses in the hope of learning lessons to avoid recurrence. Reporting radiologists and aviation operators share a dependence on humans interacting with complex technology. Larson and Nance [6] have looked at peer review in aviation and suggested how similar techniques can be applied to improve performance in radiology. They point

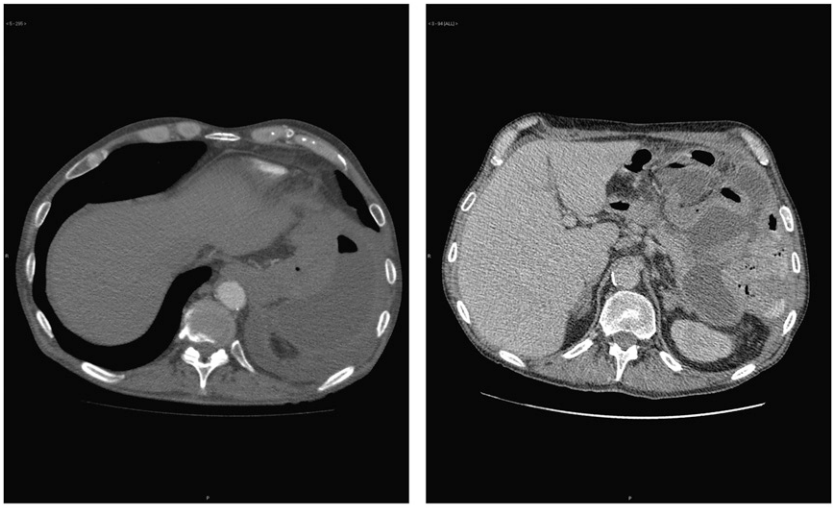
Figure 1. (a) Ultrasound of kidney reported as normal. (b) CT scan few days later shows large renal mass.



(a) (b)

out that while in the past investigations often laid blame on individual pilots, more recent trends have been to identify system failures leading to untoward incidents. This is very pertinent to radiology error peer review where individual radiologists are in the firing line for blame. Similar to aviation incidents, radiological errors are often part of a complex system failure. Larson and Nance state that "It became increasingly apparent to aviation experts by the 1970s that the underlying causes of human failures were a systemic problem and had to be treated as such". They point out that although peer review systems in radiology attempt to encourage feedback and learning, they almost always focus on quantification of error or seek to pursue quantification of error and individual performance improvement simultaneously [14–18]. Larson and Nance suggest that this is contrary to current thinking in regard to aviation safety. This echoes FitzGerald's statement from 2001 [19] that "While individuals have a duty to progressively improve their performance, the experience of safety cultures in other high-risk human activities has shown that a system approach of root cause analysis is the method required to reduce error significantly". In the United Kingdom, with the introduction of revalidation and relicensing for doctors, there is increasing pressure for quantitative assessment of a doctor's performance. The GMC has recommended such data [9] and the RCR has tools to facilitate this [20]. If quantitative assessment pertaining to individuals is to be used for such purposes, the methodology must be robust and the results reproducible. There are some commercial radiology companies using peer-review scores to monitor performance and influence continued employment, and this practice has been suggested in a peer-reviewed publication [21]. As FitzGerald has stated, not all radiological discrepancies are errors and radiological discrepancy peer review could be abused [22].

Figure 2. (a) Early phase CT pulmonary angiogram showing subphrenic collection. (b) Formal portal venous phase scan showing subphrenic collection.



(a) (b)

Figure 3. (a) Chest radiograph reported as pulmonary shadowing. (b) Subsequent CT scan showing marked lymphadenopathy.

(a) (b)

We have investigated consistency in scoring radiological errors in the peer-review process. All raters were experienced consultant radiologists and used the scoring system suggested by the RCR. We have shown a poor degree of interrater agreement. This opens questions as to whether such scores should be used in monitoring radiological performance and whether they should be considered when appraising individuals. In its present form, as used by our department, the scores seem to be of limited value.

Can we improve our scoring consistency? There are a number of steps that may help [23]

1. Change the scoring system. The distinction between the grading of the radiological error and the grading of its clinical significance may be important. Melvin et al [2] used a separate score for each. A subtle radiological miss can have serious clinical outcomes, whereas an obvious miss may have no influence on patient care. Separating these factors may reduce variation in the scoring of radiological discrepancies. This is done in some scoring systems, such as RADPEER [7] and the scoring recommended by the Faculty of Radiologists in the Republic of Ireland [24]. Disagreement has been expressed openly at South Glasgow discrepancy meetings about how much scores should reflect influence on patient outcome. The RCR in its guidance for discrepancy meetings [1] points out that judging the clinical significance of an imaging report is problematic. We feel it may be worth separating the scores in this way. An alternative approach would be to group errors into those which had clinical impact and those which did not.
2. Further classification into error type [25] may be useful to identify recurring trends but will add complexity to the process.
3. Validation of the scoring system would be ideal. The RADPEER system has been widely used and is well tested although as yet it has not been formally validated [3].
4. Clearly define the task for raters. Providing agreed standards for how scores should be applied has the potential to improve consistency.
5. Selecting raters. In our practice, discrepancy meetings are held on a departmental basis. Everyone present is encouraged to score, although for the purposes of this study scores by doctors in training were excluded. It may be that a fixed cohort of raters would give better consistency.
6. Train raters. Radiologists applying discrepancy scores could be trained in using the defined criteria in two above, and practice on a standard set of specimen cases may improve scoring agreement.
7. Remove outliers. It may be that on analysis, some raters are persistent outliers in their scores and these could be removed.
8. Introduce a process of external peer review with independent scoring.

CONCLUSION

In our practice, the interrater reliability in scoring radiological errors is poor. Great care needs to be taken in how such scores are interpreted in relation to an individual radiologist's performance. Unless such scores can be made reliable and reproducible, they should not be used as a measure of a radiologist's reporting ability. If used in appraisal or revalidation, this caveat needs to be applied and the full range of scores should be considered rather than a simple mathematical mean. We discuss possible actions that may improve scoring agreement including separating radiological error from clinical impact, also selection and training of scorers and strict definition of the scoring task. Efforts are needed to improve interrater consistency if radiological error scoring is to be a worthwhile exercise. Discrepancy meetings are an important mechanism for learning from errors, but quantification of error severity is shown to be subjective and may not be a valid exercise. If scores are applied, care must be taken that undue significance is not given to them and we support the RCR statement [1] that "discrepancy meetings cannot be used to derive error rates for individual radiologists".

REFERENCES

1. Royal College of Radiologists. Standards for radiological discrepancy meetings. London, UK: RCR; 2007.
2. 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. doi: 10.1016/j.crad.2004.01.016.
3. Jackson VP, Cushing T, Abujudeh HH, Borgstede JP, Chin KW, Grimes CK, et al. RADPEER scoring white paper. J Am Coll Radiol 2009;6;21–5. doi: 10.1016/j.jacr.2008.06.011.
4. Borgstede JP, Lewis RS, Bhargavan M, Sunshine JH. RADPEER™ quality assurance program: a multifacility study of interpretative disagreementrates. J Am Coll Radiol 2004;1:59–65. doi: 10.1016/S1546-1440(03)00002-4.
5. Larson PA, Pyatt RS Jr, Grimes CK, Abhjudeh HH, Chin KW, Roth CJ. Getting the most out of RADPEER™. J Am Coll Radiol 2011;8:543–8. doi: 10.1016/j.jacr.2010.12.018.
6. Larson DB, Nance JJ. Rethinking peer review: what aviation can teach radiology about performance improvement. Radiology 2011;259:626–32. doi: 10.1148/radiol.11102222.
7. mir-online.org [homepage on the internet]. Riga, Latvia: Unilabs Teleradiology; 2009. Available from: http://www.mir-online.org/html/img/pool/A_teleradiology_providers_approach_accepted.pdf
8. Hussain S, Hussain JS, Karam A, Vijayaraghavan G. Focused peer review: the end game of peer review. J Am Coll Radiol 2012;9:430–3. doi: 10.1016/j.jacr.2012.01.015.
9. General Medical Council. Ready for revalidation supporting information for appraisal and revalidation. London, UK: General Medical Council; 2012.
10. Fleiss JL. Statistical methods for rates and proportions. 2nd edn. New York, NY: John Wiley & Sons; 1981.
11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–74.
12. Berlin L. Radiologic errors and malpractice: a blurry distinction. AJR Am J Roentgenol 2007;189:517–22. doi: 10.2214/AJR.07.2209.
13. Berlin L. Accuracy of diagnostic procedures: has it improved over the past five decades? AJR Am J Roentgenol 2007;188:1173–8. doi: 10.2214/AJR.06.1270.
14. Mahgerefteh S, Kruskal JB, Yam CS, Blachar A, Sosna J. Peer review in diagnostic radiology: current state and a vision for the future. RadioGraphics 2009;29:122131. doi: 10.1148/rg.295095086.
15. Sheu YR, Feder E, Balsim I, Levin VF, Bleicher AG, Branstetter BF 4th. Optimizing radiology peer review: a mathematical model for selecting future cases based on prior errors. J Am Coll Radiol 2010;7:431–8. doi: 10.1016/j.jacr.2010.02.001.
16. Liu PT, Johnson CD, Miranda R, Patel MD, Phillips CJ. A reference standard-based quality assurance program for radiology. J Am Coll Radiol 2010;7:61–6. doi: 10.1016/j.jacr.2009.08.016.
17. Steele JR, Hovsepian DM, Schomer DF. The joint commission practice performance evaluation: a primer for radiologists. J Am Coll Radiol 2010;7:425–30. doi: 10.1016/j.jacr.2010.01.027.
18. Abujudeh HH, Boland GW, Kaewlai R, Rabiner P, Halpern EF, Gazelle GS, et al. Abdominal and pelvic computed tomography(CT) interpretation: discrepancy rates among experienced radiologists. Eur Radiol 2010;20:1952–7. doi: 10.1007/s00330-010-1763-1.
19. FitzGerald R. Error in radiology. Clin Radiol 2001;56:938–46. doi: 10.1053/crad.2001.0858.
20. The Royal College of Radiologists. Specialty standards and supporting information for revalidation. London, UK: The Royal College of Radiologists; 2010.
21. Hussain S, Hussain JS, Karam A, Vijayaraghavan G. Focused peer review: the end game of peer review. J Am Coll Radiol 2012;9:430–3.e1. doi: 10.1016/j.jacr.2012.01.015.
22. FitzGerald R. Radiological error: analysis, standard setting, targeted instruction and teamworking. Eur Radiol 2005;15:1760–7. doi: 10.1007/s00330-005-2662-8.
23. Gwet KM, ed. Handbook of inter-rater reliability: the definitive guide to measuring the extent of agreement among multiple raters. 3rd edn. Gaithersburg, MD: Advanced Analytics, LLC; 2012.
24. Faculty of Radiologists. Guidelines for the implementation of a national quality assurance programme in radiology. Version 2.0. Dublin, Ireland: Faculty of Radiologists; 2011.
25. Renfrew RL, Franken EA, Berbaum KS, Weigelt FH, Abu-Yousef MM. Error in radiology: classification and lessons in 182 cases presented at a problem case conference. Radiology 1992;183;145–50.