AJNR

Interobserver Agreement in the Interpretation of Outpatient Head CT Scans in an Academic Neuroradiology Practice

G. Guérin, S. Jamali, C.A. Soto, F. Guilbert and J. Raymond

AJNR Am J Neuroradiol 2015, 36 (1) 24-29
doi: https://doi.org/10.3174/ajnr.A4058
http://www.ajnr.org/content/36/1/24

Interobserver Agreement in the Interpretation of Outpatient Head CT Scans in an Academic Neuroradiology Practice

G. Guérin, S. Jamali, C.A. Soto, F. Guilbert, and J. Raymond

ABSTRACT

BACKGROUND AND PURPOSE: The repeatability of head CT interpretations may be studied in different contexts: in peer-review quality assurance interventions or in interobserver agreement studies. We assessed the agreement between double-blind reports of outpatient CT scans in a routine academic practice.

MATERIALS AND METHODS: Outpatient head CT scans (119 patients) were randomly selected to be read twice in a blinded fashion by 8 neuroradiologists practicing in an academic institution during 1 year. Nonstandardized reports were analyzed to extract 4 items (answer to the clinical question, major findings, incidental findings, recommendations for further investigations) from each report, to identify agreement or discrepancies (classified as class 2 [mentioned or not mentioned or contradictions between reports], class 1 [mentioned in both reports but diverging in location or severity], 0 [concordant], or not applicable), according to a standardized data-extraction form. Agreement regarding the presence or absence of clinically significant or incidental findings was studied with κ statistics.

RESULTS: The interobserver agreement regarding head CT studies with positive and negative results for clinically pertinent findings was 0.86 (0.77–0.95), but concordance was only 75.6% (67.2%–82.5%). Class 2 discrepancy was found in 15.1%; class 1 discrepancy, in 9.2% of cases. The κ value for reporting incidental findings was 0.59 (0.45–0.74), with class 2 discrepancy in 29.4% of cases. Most discrepancies did not impact the clinical management of patients.

CONCLUSIONS: Discrepancies in double-blind interpretations of head CT examinations were more common than reported in peer-review quality assurance programs.

ABBREVIATION: CHUM = Centre Hospitalier de l'Université de Montréal

The delivery of optimal radiology services may require continuous vigilance and perhaps quality assurance interventions.[1-3] The content of these interventions may not be evident, however. In addition, the manner in which the error, discrepancy, and disagreement should be handled both in theory and in clinical practice is evolving.[4]

Discrepancies in peer-review approaches have been known for a long time.[5-7] In 1959, Garland[8] claimed that radiologists missed approximately 30% of tuberculosis cases in screening chest x-ray examinations.[9] Garland's report launched a series of investigations that continue today. However, there is no consensus on a standard method or protocol for evaluating errors and discrepancies in imaging reports, and rates published in the literature differ widely.[1-3,10-14] Multiple variations in study parameters, including sampling sources, methods, imaging modalities, specialties, categories, interpreter training levels, and degrees of blinding, may have contributed to this wide spectrum.[2,3,9]

Recently, CT and MR imaging reports of the head, neck, and spine were re-read by staff neuroradiologists, and a 2% clinically significant discrepancy rate was found, an excellent result compared with the 3%–6% radiologic error rates published in general radiology practices.[3,15,16]

To anyone who has studied reliability or precision of diagnostic imaging tests, such levels of disagreement between interpretations may appear unbelievably low. Peer-review quality assurance "errors and discrepancies" and disagreements in reliability studies of imaging test interpretations may not measure the same things. Discrepancies in the reporting of imaging studies can thus be approached from at least 2 different perspectives.

Received May 9, 2014; accepted after revision June 12.

From the Department of Radiology (G.G., C.A.S., F.G., J.R.), Centre Hospitalier de l'Université de Montréal, Notre-Dame Hospital, Montreal, Quebec, Canada; and Laboratory of Interventional Neuroradiology (S.J., J.R.), Centre Hospitalier de l'Université de Montréal, Notre-Dame Hospital Research Centre, Montreal, Quebec, Canada.

Please address correspondence to Jean Raymond, MD, CHUM–Notre-Dame Hospital, Interventional Neuroradiology, 1560 Sherbrooke East, Pavillion Simard, Room Z12909, Montreal, Quebec, Canada H2L 4M1; e-mail: jean.raymond@umontreal.ca

http://dx.doi.org/10.3174/ajnr.A4058

From a quality assurance point of view, optimal radiology services require continuous quality assurance interventions. One report is the true right one, and discrepancies are errors that must be minimized. Performance can be measured; deviations and outliers can be identified, and appropriate measures can then be taken to improve performance.[1-3]

A different vocabulary is used when discrepancies are examined from a scientific point of view. In the typical absence of a criterion standard of "truth," the uncertainty is a reality that must be admitted and taken into account when using imaging reports for clinical decisions. Reliability and agreement can be measured by using proper methods, including independent readings; and concordant or diverging verdicts can be tabulated and summarized, though imperfectly, by using marginal sums and appropriate statistical tools (such as κ statistics). No test and certainly no imaging study requiring an element of interpretation will ever be perfectly repeatable.

Reconciliation between these 2 perspectives is desirable. The credibility of quality assurance programs disconnected from scientific methods is shaky. If only errors could be defined, perhaps as discrepancies beyond "normal discrepancies." Unfortunately attempts to define an acceptable level of radiologic discrepancy are probably futile. Multiple variables are at play, and distinctions, even between acceptable discrepancy and negligence, may remain blurry.[17]

To our knowledge, reliability and agreement in the independent interpretation of head CT scans by expert neuroradiologists in a routine academic clinical practice have not been reported. In contrast to a peer-review approach, examining discrepancies after independent interpretations of clinical cases in everyday practice and looking for consensus on discrepant cases may provide a realistic and favorable framework for continuous quality improvement for each and all professionals, rather than the identification of specific deviant individuals. With this end in view, we studied the discrepancy in independent double readings of outpatient head CT scans in an academic practice. We hypothesized that our study would show a discrepancy rate in the range of $\geq$5%.

MATERIALS AND METHODS

The present article was written in accordance with the "Guidelines for Reporting Reliability and Agreement Studies" framework.[18] A protocol was initially prepared, including a detailed data-collection form for each interpretation, prespecified plans for comparing pairs of reports, and statistical analyses. The protocol was discussed and accepted by all participating radiologists and by the head of the clinical neuroradiology service. Readers were informed that the identity of the participating radiologists would remain anonymous. As part of a pilot quality-improvement program, the necessity for obtaining informed consent was waived. Nonurgent outpatient head CT scans were interpreted on a double-blind basis within 2 consecutive days. The 2 reports were analyzed by 1 author (G.G.), who filled out the corresponding data-collection forms.

Patients

During 12 months (between July 2012 and July 2013), 119 outpatients (71 women; 48 men; mean age, 60.5 $\pm$ 15.4 years) with head CT requests from any outpatient clinic, including a neuro-oncology clinic (n = 35, 29.4%), were randomly selected from the Centre Hospitalier de l'Université de Montréal (CHUM), Notre-Dame Hospital imaging library on a basis of 2–5 cases per week, on certain weeks only (cases could only be submitted to double reading when the first adjudicator was available to review reports in a timely fashion). No formal sample-size calculations were performed because this pilot project was considered exploratory, but a statistical consultant advised that >100 patients were necessary to provide meaningful confidence intervals.

Cases were blindly and randomly selected and sent back, just as a "new case" would be interpreted within the workflow of the next consecutive day, for an independent re-reading. Cases that, by chance, happened to be read twice by the same reader were excluded from analyses (n = 5). CT scans from the emergency department or performed on hospitalized patients were excluded for methodologic and ethical considerations. The first report was automatically erased from the patient file when the second report became available but was saved in a separate file for this study. In this fashion, the second reader never had access to the first report, and head CTs were interpreted twice in a blinded fashion. All examinations were anonymously integrated in the daily workload, and both reports were dictated within 2 consecutive days. The second readings were made available to clinicians and became the permanent official report, unless flagged by G.G. When a finding was mentioned in the first but not in the second report, the second reader was asked to review or discuss the case with the first reader and amend the report if necessary. When the most inclusive or "safest" report was the second official one, no immediate revision was undertaken.

Readers

The readings were performed by 8 certified staff radiologists, fellowship-trained and specialized in neuroradiology with clinical experience ranging from 1 to 37 years (mean, 17.9 $\pm$ 7.3 years; median, 17 years). They worked independently, rotating each day within the same tertiary institution. They were aware that a study was ongoing but could not guess which patients were included (<1% of the patients being examined were part of the study).

Data Extraction and Analyses

Each report was analyzed by 1 author (G.G.) to extract the following 4 items: 1) response to the question raised by the requesting physician, 2) major/clinically pertinent findings (defined as related to the clinical question or requiring immediate management, such as recent ischemic lesions, tumor evolution or recurrence, acute sinusitis, and so forth), 3) incidental findings (defined as not being related to the clinical question and requiring no immediate management, such as chronic sinusitis, cerebral atrophy, old lacunar infarcts), 4) further investigations being proposed.

Pairs of reports were compared, and agreement was assessed regarding the following: 1) detection agreement or agreement on the presence or absence of an abnormal finding, sorting out positive-versus-negative test results; and 2) agreement on the description of the findings regarding the 4 previously mentioned items. For each of the first 3 items (clinical question, major clinically pertinent finding, and incidental finding), concordance was rated according to the content of the reports (and not according to the

Table 1: Agreement among double-blind reports

	Concordant Readings (CI 95%)	Class 2 Discrepancies	Class 1 Discrepancies
Clinically pertinent findings	75.6% ($n = 90$) (67.2%–82.5%)	15.1% ($n = 18$)	9.2% ($n = 11$)
Incidental findings	65.5% ($n = 78$) (56.6%–73.5%)	29.4% ($n = 35$)	5% ($n = 6$)
Answer to the clinician's question	86.7% ($n = 85$) (78%–92.5%)	12.2% ($n = 12$)	1% ($n = 1$)
Further investigations	86.6% ($n = 103$) (79.2%–91.6%)	13.4% ($n = 16$)	

Table 2: Class 2 discrepancies in clinically significant findings

	Age (yr)	Sex	NC or CE	Context	Discrepancies
1	66	F	CE	F-up meningioma	Progression vs stable
2	82	F	CE	F-up meningioma	Progression vs stable
3	47	M	NC	F-up subdural hematoma	Complete resolution vs partial regression
4	60	F	NC	F-up hydrocephalus	Stable vs progression
5	47	F	CE	F-up metastasis resection	No recurrence vs recurrence suspicion
6	77	M	CE	Lymphoma	Normal vs clival lesion
7	76	M	NC	F-up pituitary apoplexy	Unchanged vs new hemorrhage
8	63	M	NC	F-up glaucoma	Optic nerve atrophy vs normal
9	66	M	NC	F-up subdural hematoma	Persistence vs complete resolution
10	64	F	NC	Post-op adenoma	Residual mass vs normal
11	75	F	CE	Seizures	Schizencephaly vs no mention
12	74	M	NC	F-up subdural hematoma	Rebleeding vs no rebleeding
13	58	F	CE	F-up metastases	No recurrence vs recurrence
14	37	F	NC	F-up glioblastoma	Herniation vs no mention
15	62	F	CE	Breast CA	Choroidal mets suspicion vs no mention
16	77	F	NC	Dementia	White matter disease vs no mention
17	76	M	CE	Memory loss	Cerebral trophy vs no mention
18	42	M	CE	F-up glioblastoma	Progression vs stable

Note:—NC indicates noncontrast CT; CE, contrast-enhanced CT; CA, cancer; F-up, follow-up; Post-op, postoperative; mets, metastases.

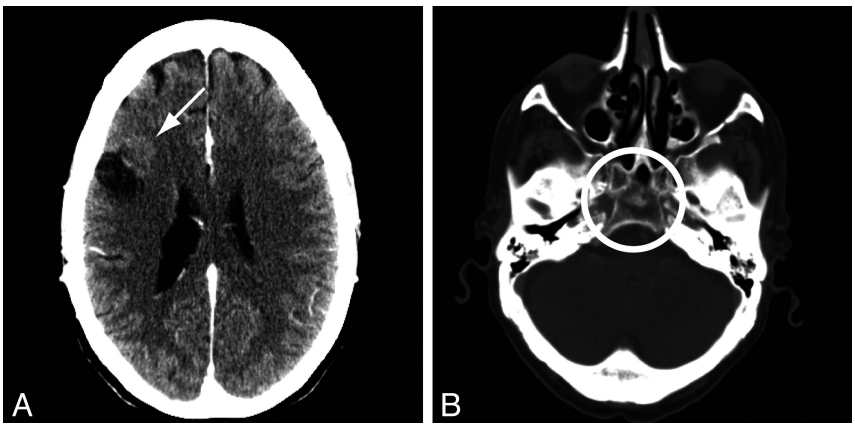

FIG 1. Class 2 discrepancies in pertinent findings. *A*, A patient followed for recurrence of a resected breast cancer metastasis; one reader reported a recurrence (*arrow*), whereas the other observer did not mention this finding. *B*, A lytic lesion of the clivus (*circle*) was reported by one but not the other reader.

clinical significance of the discordance) as the following: not applicable (in the absence of a clinical question or of any finding); 0, concordant; class 1 discordance, the same findings noted in both reports but in different locations or diverging in severity; class 2 discordance, the finding was not mentioned in one of the reports or opinions diverged on the evolution or recurrence of a lesion.

Subsequently, a senior author (J.R.) who had not previously read any of the examinations reviewed all reports and data-collection forms (unmasked).

Statistical Analyses

The concordance rates for responses to the clinical questions raised by the referring physicians, clinically pertinent findings, incidental findings, and proposed investigations were tabulated to provide proportions (with 95% confidence intervals). Class 2 discrepancies in reporting pertinent clinical findings from contrast-enhanced and nonenhanced CTs were compared by using the Fisher exact test. The Cohen κ was calculated (with 95% confidence intervals) for the presence/absence of clinically pertinent findings and incidental findings. κ values were interpreted according to Landis and Koch.[19]

RESULTS

There were 119 CT examinations; 53 (44.5%) were contrast-enhanced. A clinical question was formulated by the requesting physician in 98 cases (82.4%) and the reader responses were concordant in 88.1% of cases (95% CI, 80.6%–92.9%). The Cohen κ for positive studies was 0.86 (0.77–0.95) and 0.59 (0.45–0.74) for the presence/absence of any clinically pertinent findings and for incidental findings, respectively.

Readings were in agreement for 75.6% (67.2%–82.5%), 65.5% (56.6%–73.5%), and 86.6% (79.2%–91.6%) of clinically pertinent findings, incidental findings, and recommendations for further investigations, respectively (Table 1). Rates were similar for the 2 subgroups (referred from the neuro-oncology clinic or from all other clinics).

Class 2 discrepancies in reporting clinically pertinent findings were found in 18 cases (15.1%). They are summarized in Table 2. Examples include the presence or absence of a tumor recurrence

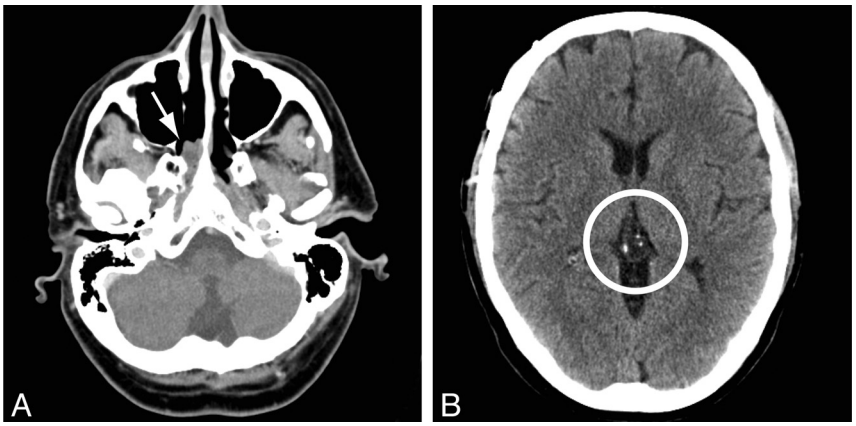

FIG 2. Class 2 discrepancies in incidental findings. *A*, One reader reported a polypoid posterior nasal lesion (*arrow*), whereas the other observer did not mention this incidental finding. *B*, A pineal cyst (*circle*) was mentioned in only 1 of the 2 reports.

($n = 4$, Fig 1*A*), the growth of a meningioma ($n = 2$), the evolution of chronic subdural hematomas ($n = 2$), and the presence of a lytic bone lesion (Fig 1*B*). These discrepancies were normally distributed between readers ($n = 1, 3, 5, 7, 11, 5, 3, 1$ for 36 discrepant reports). There was no significant difference between contrast-enhanced ($n = 9$ of 53) and nonenhanced studies ($n = 9$ of 66, $P = .62$).

Class 1 discrepancies in clinically pertinent findings were seen in the interpretation of 11 cases. Examples include the location of recent ischemic lesions ($n = 2$), tumor extensions ($n = 2$), or the disconnection of a ventricular shunt ($n = 1$).

Class 2 discrepancies in reporting incidental findings were seen in 35 cases. Examples include the presence or absence of white matter disease ($n = 10$), chronic sinusitis ($n = 6$), old strokes ($n = 5$), atrophy ($n = 4$), a nasal polyp (Fig 2*A*), or a pineal cyst (Fig 2*B*). Class 1 discrepancies in calling incidental findings were seen in 6 cases. Examples include the location of lacunar infarcts ($n = 2$) or the extension of chronic sinusitis ($n = 2$).

The senior author confirmed the discrepancies identified by the first adjudicator in all cases, except for 2 minor modifications in the categorization of incidental findings.

DISCUSSION

The salient findings of this study are the following: 1) The interobserver agreement regarding head CT outpatient studies with positive and negative findings measured by the Cohen κ was 0.86 (almost perfect), but 2) class 2 discrepancies in clinically pertinent findings were still found in 15% of cases, above the 5% level we expected from the 3%–6% discrepancy rates reported in previously published neuroradiologic peer-review studies.[3,15,16]

Different aspects of study design may explain our results. The definitions we used were somewhat arbitrary, though they were inspired from similar studies.[3,16] Perhaps they were rigorously applied during data extraction. In the absence of standardization of reporting, some variability in the attribution of categoric verdicts to the content of reports of different styles is inevitable.

The levels of disagreement that we observed are not unheard of. Robinson et al[20] investigated the concordance among 3 independent observers. Their study showed agreement in 51%, 61%, and 74% of abdominal, chest, and skeletal x-rays, respectively. They also assessed performance by calculating κ statistics of interobserver agreement. Weighted κ values between pairs of observers were higher with skeletal (0.76–0.77) than with chest (0.63–0.68) or abdominal (0.50–0.78) examinations. In a meta-analysis conducted by Wu et al,[2] the global discrepancy rate was 7.7% (including a major discrepancy rate of 2.4%). The major discrepancy rate varied according to body region: It was lower for head (0.8%) and spine CT (0.7%) than chest (2.8%) and abdominal CT (2.6%). Blinding of the reference radiologist to the initial report was, however, associated with higher discrepancy rates compared with studies that lacked blinding: Not blinding the initial report yielded a much lower major discrepancy rate (2.0%; 95% CI, 1.4%–2.7%) than with blinding (12.1%; 95% CI, 4.4%–29.4%).

Hence, we believe that the main reason for the difference between our results and those of previous studies on discrepancies in head CT reports is that we used double-blind reporting. A possible explanation is that knowledge of the initial report may lead to a "satisfaction of search" error that reduces discrepancies.[2]

Most quality assurance studies published in the literature have assessed the discrepancy rates found through a peer-review approach.[2,3,6,9,21] In that context, the opinion of the second observer is considered the criterion standard, and discrepancies are meant to be errors. However, in many cases, the second radiologist worked in a setting that differs from the normal clinical context. The second interpretation can be biased in many ways, by knowledge of the first interpretation, knowledge of the identity of the first interpreter, or even knowledge of the clinical evolution and outcome of the patient (ie, hindsight bias). Of course, the second interpreter is also aware that he or she is working within an audit process.[3] Other studies that have used a double-blind method in a much larger number of patients have found lower discrepancy rates than those reported here.[1] In this case, we suspect that the medical director or quality assurance radiologist knowingly working within an audit process may have biased, according to a principle of charity, the selective identification of the most serious discrepancies only. Our study being done in an anonymous, nonblaming context, with no intent to identify deviant individuals differs from many quality assurance studies and may explain why more frequent disagreements could be identified. Fortunately, most discrepancies had minimal clinical impact in terms of patient management or outcome.

Discrepancies in the present study do not necessarily mean errors. The use of clinical judgment, in assessing the impact of calling a suspicious-but-uncertain finding or the pertinence of reporting an incidental finding, for example, may lead to variations from one individual report to the other.

Some societies have proposed guidelines for peer-review

methods or for holding discrepancy meetings.[22,23] Many questions remain unanswered, such as which method ought to be followed or which is most efficient.[22,23] A survey of practices in the United Kingdom has recently shown wide variation in discrepancy meetings.[24] Larson and Nance[6] have emphasized that "peer review can either serve as a coach or a judge, but it cannot do both well." There is a recent trend to shift away from the identification of deviant individuals to the overall improvement of systems and patient care, to move away from a blame culture to one of continuous feedback, learning from each other's methods of working and reporting and from each other's mistakes.

The work presented here was a pilot project. It was not designed to compete with commercially available peer-review systems. The process requires the active participation of a dedicated third physician (in addition to the 2 readers) and is time-consuming. In this particular case, the third physician was a senior resident, who considered the experience informative and enriching. Perhaps this pilot project could shed some light on how we should understand and design methods purported to improve patient care, including systematic quality assurance peer-review systems or discrepancy meetings. Neuroradiologists who participated in this project were initially surprised to see that discrepancies were common. They have started to exchange opinions on the pertinence of incidental findings, for example. The sharing of practices in an open-minded learning context could be a modest but meaningful step, from solo practices to a richer, more stimulating professional experience, perhaps more favorable to better radiology services.

Limitations

There were many limitations to this study: The number of cases was small, the heterogeneity of clinical presentations was modest, and the technique was restricted to head CT. Not all patients were represented, for inpatients and cases from the emergency department were excluded. Participating radiologists were from a single, tertiary referral center. The population we studied had a high prevalence of positive findings. We were careful to explicitly define categories in advance in the research protocol and to make judgments as objective as possible, but such categorization as "incidental finding" remains a judgment that may depend on context, interpretation, local culture, and training. The data were extracted by a single individual, and the review of the data extraction forms by the second rater was not blinded. We had made that decision at the time of study design: A blinded second adjudication could have produced variability in categorization and discrepancies between adjudicators, at the risk of launching an infinite regress. These limitations may have contributed to overestimating agreement between reports. Yet, discrepancies were more frequent than those in previous, nonblinded, peer-review reports, a finding that remains to be confirmed by other double-blind studies.

Comparing reports that widely vary in format and style is fastidious and time-consuming. Some standardization of reporting may be necessary to ease the identification of discrepancies. Adopting a rule of keeping the most inclusive report as the official report, combined with double-blind reading, may increase the number of "overcalls," compared with a single reading. Finally, no attempt was made to determine the real diagnoses, for the study was focused on agreement and not on diagnostic accuracy. We did not try to study the causes of discrepancies. The protocol neither included a clinical follow-up period, to monitor what actually happened to study patients, nor tried to evaluate the theoretic impact of any report on clinical decisions.

CONCLUSIONS

Double-blind reading of head CT scans can show class 2 discrepancies in 15% of cases. If duplicated, this finding should be taken into account when planning quality assurance interventions.

ACKNOWLEDGMENTS

The authors acknowledge the participation of the attending neuroradiologists of CHUM and the contribution of the PACS and imaging library staff.

Disclosures: Jean Raymond—*UNRELATED*: *Grants/Grants Pending*: Canadian Institutes of Health Research,* *Comments*: funding for a clinical trial (Canadian UnRuptured aneurysm Endovascular versus Surgery [CURES]).* *Money paid to the institution.

REFERENCES

1. Soffa DJ, Lewis RS, Sunshine JH, et al. **Disagreement in interpretation: a method for the development of benchmarks for quality assurance in imaging.** *J Am Coll Radiol* 2004;1:212–17
2. Wu MZ, McInnes MD, Blair Macdonald D, et al. **CT in adults: systematic review and meta-analysis of interpretation discrepancy rates.** *Radiology* 2014;270:717–35
3. Babiarz LS, Yousem DM. **Quality control in neuroradiology: discrepancies in image interpretation among academic neuroradiologists.** *AJNR Am J Neuroradiol* 2012;33:37–42
4. McCoubrie P, FitzGerald R. **Commentary on discrepancies in discrepancy meetings.** *Clin Radiol* 2014;69:11–12
5. Bender LC, Linnau KF, Meier EN, et al. **Interrater agreement in the evaluation of discrepant imaging findings with the Radpeer system.** *AJR Am J Roentgenol* 2012;199:1320–27
6. Larson DB, Nance JJ. **Rethinking peer review: what aviation can teach radiology about performance improvement.** *Radiology* 2011;259:626–32
7. Mucci B, Murray H, Downie A, et al. **Interrater variation in scoring radiological discrepancies.** *Br J Radiol* 2013;86:20130245
8. Garland LH. **Studies on the accuracy of diagnostic procedures.** *Am J Roentgenol Radium Ther Nucl Med* 1959;82:25–38
9. Berlin L. **Accuracy of diagnostic procedures: has it improved over the past five decades?** *AJR Am J Roentgenol* 2007;188:1173–78
10. Abujudeh HH, Boland GW, Kaewlai R, et al. **Abdominal and pelvic computed tomography (CT) interpretation: discrepancy rates among experienced radiologists.** *Eur Radiol* 2010;20:1952–57
11. Agrawal A, Pandit M, Kalyanpur A. **Systematic survey of discrepancy rates in an international teleradiology service.** *Emerg Radiol* 2011;18:23–29
12. Chung JH, Strigel RM, Chew AR, et al. **Overnight resident interpretation of torso CT at a level 1 trauma center an analysis and review of the literature.** *Acad Radiol* 2009;16:1155–60
13. Goddard P, Leslie A, Jones A, et al. **Error in radiology.** *Br J Radiol* 2001;74:949–51
14. Loevner LA, Sonners AI, Schulman BJ, et al. **Reinterpretation of cross-sectional images in patients with head and neck cancer in the setting of a multidisciplinary cancer center.** *AJNR Am J Neuroradiol* 2002;23:1622–26
15. Borgstede JP, Lewis RS, Bhargavan M, et al. **RADPEER quality assurance program: a multifacility study of interpretive disagreement rates.** *J Am Coll Radiol* 2004;1:59–65
16. Siegle RL, Baram EM, Reuter SR, et al. **Rates of disagreement in imaging interpretation in a group of community hospitals.** *Acad Radiol* 1998;5:148–54

17. Berlin L. **Radiologic errors and malpractice: a blurry distinction.** *AJR Am J Roentgenol* 2007;189:517–22
18. Kottner J, Audige L, Brorson S, et al. **Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed.** *J Clin Epidemiol* 2011;64:96–106
19. Landis JR, Koch GG. **The measurement of observer agreement for categorical data.** *Biometrics* 1977;33:159–74
20. Robinson PJ, Wilson D, Coral A, et al. **Variation between experienced observers in the interpretation of accident and emergency radiographs.** *Br J Radiol* 1999;72:323–30
21. Herman PG, Gerson DE, Hessel SJ, et al. **Disagreements in chest roentgen interpretation.** *Chest* 1975;68:278–82
22. O'Keeffe MM, Piche S, Mason AC. **The Canadian Association of Radiologists (CAR) Guide to Peer Review Systems.** Approved September 10, 2011; Amended July 16, 2012. http://www.car.ca/uploads/standards%20guidelines/20120831_en_peer-review.pdf. Accessed November 21, 2013
23. The Royal College of Radiologists. **Standards for radiology discrepancy meetings.** London ETRCoR, 2007. https://www.rcr.ac.uk/docs/radiology/pdf/Stand_radiol_discrepancy.pdf. Accessed November 21, 2013
24. Prowse SJ, Pinkey B, Etherington R. **Discrepancies in discrepancy meetings: results of the UK national discrepancy meeting survey.** *Clin Radiol* 2014;69:18–22