AJNR

Quality Control in Neuroradiology: Discrepancies in Image Interpretation among Academic Neuroradiologists

L.S. Babiarz and D.M. Yousem

AJNR Am J Neuroradiol published online 27 October 2011
http://www.ajnr.org/content/early/2011/10/27/ajnr.A2704

Published October 27, 2011 as 10.3174/ajnr.A2704

PRACTICE PERSPECTIVES

Quality Control in Neuroradiology: Discrepancies in Image Interpretation among Academic Neuroradiologists

L.S. Babiarz
D.M. Yousem

SUMMARY: Prior studies have found a 3%–6% clinically significant error rate in radiology practice. We set out to assess discrepancy rates between subspecialty-trained university-based neuroradiologists. Over 17 months, university neuroradiologists randomly reviewed 1000 studies and reports of previously read examinations of patients in whom follow-up studies were read. The discrepancies between the original and "second opinion" reports were scored according to a 5-point scale: 1, no change; 2, clinically insignificant detection discrepancy; 3, clinically insignificant interpretation discrepancy; 4, clinically significant detection discrepancy; and 5, clinically significant interpretation discrepancy. Of the 1000 studies, 876 (87.6%) showed agreements with the original report. The neuroradiology division had a 2.0% (20/1000; 95% CI, 1.1%–2.9%) rate of clinically significant discrepancies involving 8 CTs and 12 MR images. Discrepancies were classified as vascular ($n = 7$), neoplastic ($n = 9$), congenital ($n = 2$), and artifacts ($n = 2$). Individual neuroradiologist's scores ranged from 0% to 7.7% ± 2.3% ($n = 18$). Both CT and MR imaging studies had a discrepancy rate of 2.0%. Our quality assessment study could serve as initial data before intervention as part of a PQI project.

ABBREVIATIONS ABR = American Board of Radiology; ACGME = Accreditation Council for Graduate Medical Education; CI = confidence interval; PQI = practice quality improvement

Radiologic detection and interpretation errors will not be fully eliminated until the advent of "perfect diagnostic tests" and "perfect observers."[1,2] In the meantime, radiologists, similar to other physicians, struggle with assessing physician performance and reporting quality, to improve and deliver the best care possible.[3] L. Henry Garland pioneered the work on radiologic errors more than 60 years ago.[4-6] He uncovered a 30% rate of missed radiologic findings in a series of radiographs with abnormal findings among expert reviewers. Subsequently Garland's results have been replicated by other researchers.[7-9] Most interesting, comparable rates of "mistakes" were discovered in other specialties.[10,11] In deriving the radiologic error rate, Garland used exclusively abnormal studies—that is, he tested radiologists in environments in which disease prevalence reached 100%. Because in the typical clinical setting, there are a substantial number of examinations with normal findings, Garland hypothesized that the expected radiologic error rate in everyday practice is closer to 5%.

Subsequent studies confirmed the radiologic error rate in all-comers radiology practice to be in the 3%–6% range.[3,12-14] Soffa et al[14] sampled approximately 7000 cases read by 26 radiologists and uncovered a 3% disagreement rate in general radiology, 3.6% in diagnostic mammography, 5.8% in screening mammography, and 4.1% in sonography, yielding the overall error rate of 3.5%. Robinson et al[13] compared reports for skeletal, chest, and abdominal radiographs completed by 3 radiologists and found a 3%–6% average error rate per observer. Siegle et al[12] reviewed radiologic studies performed during a 7-year period in 6 community hospitals, including general radiology, nuclear medicine, CT, and MR imaging, and calculated a mean rate of disagreement of 4.4%. In agreement with Garland's results, practices with arguably higher disease prevalence, like academic medical centers, tend to have higher error rates (≤1.5% has been reported) than practices with a greater number of normal findings on studies, like community hospitals.[3]

In this study, we set out to compare finalized dictations on record, with second-opinion reviews of neuroradiology cases read by subspecialty-trained neuroradiologists on staff at a university hospital. We did so in an effort to assess the detection or interpretation discrepancy rate for clinically significant radiographic findings between equally qualified subspecialty readers. We hypothesized that at a major academic medical center with a large tertiary care patient population base, such discrepancy rates would be <5%. The collection of data also served as a baseline measurement before instituting a PQI initiative for improving consistent and accurate interpretations.

Materials and Methods

Data Collection

In accordance with the Health Insurance Portability and Accountability Act, our institutional review board reviewed the protocol for this retrospective study and waived the requirement for informed consent.

During 17 months (January 2009 to May 2010) as a part of this study as well as a part of a quality assurance initiative in our department, staff neuroradiologists reviewed previously read neuroradiology studies. For the first 2 current studies of the day that had prior examinations, each neuroradiologist was instructed to review and grade the most recent companion case study and report as a part of their regular workflow. The second-opinion interpretations were compared with the original reports, and the discrepancies between

Received February 26, 2011; accepted after revision April 27.

From the Russell H. Morgan Department of Radiology and Radiological Sciences, The Johns Hopkins Medical Institutions, Baltimore, Maryland.

Paper previously presented at: Annual Meeting of the American Society of Neuroradiology, June 4–9, 2011, Seattle, Washington.

Please address correspondence to D.M. Yousem, MD, Russell H. Morgan Department of Radiology and Radiological Sciences, The Johns Hopkins Medical Institutions, 600 N Wolfe St, Phipps B-100F, Baltimore, MD 21287; e-mail: dyousem1@jhu.edu

http://dx.doi.org/10.3174/ajnr.A2704

Table 1: Five-point scale used in scoring discrepancies between the original report and the second opinion reading of a study

Point Scale	Verbal Descriptor
1	No change
2	Detection discrepancy, not clinically significant
3	Interpretation discrepancy, not clinically significant
4	Detection discrepancy, clinically significant
5	Interpretation discrepancy, clinically significant

the 2 documents were scored according to a previously validated[15] 5-point rating scale. The 5-point scale allowed the following categories: 1, no change in the reading; 2, finding of a clinically insignificant detection discrepancy (eg, a missed case of mild chronic sinusitis); 3, a finding of a clinically insignificant interpretation discrepancy (eg, interpretation of an oligodendroglioma as an astrocytoma); 4, a finding of a clinically significant detection discrepancy (eg, a missed tumor); and 5, a finding of a clinically significant interpretation discrepancy (eg, interpreting a tumor as a stroke) (Table 1). In addition to the discrepancy score, staff also noted the type of imaging study of each examined case. In the event of a discrepancy (grades 2–5), staff recorded the source and nature of the discrepancy.

From the pool of the reviewed cases, 100 were randomly selected and their reports were analyzed for prevalence and type of disease.

Eleven subspecialty-certified or subspecialty-certification-eligible neuroradiologists assigned to read current cases were included in the sample and reviewed cases and reports originally read by 18 subspecialty-certified or subspecialty-certification-eligible neuroradiologists. A neuroradiologist is subspecialty-certified or subspecialty-certification-eligible following completion of a 4-year ACGME-accredited diagnostic radiology residency program and a 1-year ACGME-accredited diagnostic neuroradiology fellowship. Neuroradiologists must practice neuroradiology for 1 year after their fellowship year and then pass a 4-hour cognitive test of neuroradiology knowledge, proctored by members of the ABR, to become subspecialty-certified and receive a Certificate of Added Qualification.

Each neuroradiologist was responsible for covering the clinical service an average of 3–4 days a week. On some services, there may not have been 2 cases to review that had comparison studies because of low volumes (eg, myelography service, teleradiology service, and so forth). The 11 neuroradiologist reviewers were not allowed to review their own previous radiology reports and were instructed to skip such cases. The readers ranged in experience from 1 to 28 years post-neuroradiology fellowship training.

All studies with clinically significant discrepancies (scores of 4 and 5) were assessed by a third independent expert reviewer with 22 years of experience in neuroradiology (D.M.Y.) who had previously published data to validate the 5-point scoring system.[15] Because D.M.Y. was 1 of the 18 neuroradiologists whose reports were reviewed in this study, no adjudicating was done on the cases that he had originally read.

Following the completion of this study, each staff member was presented with a score card that showed their performance (scores of 1, 2, and 3 versus scores of 4 and 5) in absolute percentages as well as relative to the divisional average. The neuroradiologists understood that the scoring would be used as part of baseline data for a maintenance-of-certification PQI initiative.

Statistical Analysis

Total counts of all scores and their relative percentages were tabulated for each staff member as well as for the whole neuroradiology division. For the purpose of calculating the overall and individual discrepancy rates, scores of 1, 2, and 3 (clinically insignificant findings), and scores of 4 and 5 (clinically significant findings) were collated into 2 separate groups and the 2 groups were compared. The mean, 95% CI, SD, minimum, and maximum for the overall and individual discrepancy rates were computed. Discrepancy rates were also calculated for each imaging technique (CT versus MR imaging) and body region imaged (brain/head and neck versus spine versus bony structures) and then were compared by using a standard *t* test. Studies of the brain and head and neck for CT and for MR imaging examinations were collated into the brain/head and neck category for the purpose of the body region analysis due to the very low number of head and neck studies (<5%) in our sample. Additionally, years of experience and study volumes were compared between the reviewers with and without significantly discrepant scores (scores of 4 or 5 were deemed significant). For years of experience and discrepancy scores, the Spearman rank correlation coefficient was also computed. Finally, given that a large portion of the neuroradiologists were trained at the same home institution and thus may have had similar interpretation tendencies and biases, we also looked at the difference in discrepancy rates between the neuroradiologists trained at the home and at other institutions. In all analyses, significant differences were rated as $P < .05$.

Table 2: Final distribution of discrepancy scores

Score	No. of Studies	% Total
1	876	87.6
2	75	7.5
3	29	2.9
4	14	1.4
5	6	0.6
1, 2, and 3	980	98.0
4 and 5	20	2.0
Total	1000	100.0

Results

One thousand neuroradiology studies originally read by 18 neuroradiologists (present and past) were reread by the 11 neuroradiologists currently on staff. On average, 55.6 studies were reviewed for each of the 18 original readers (SD, 40.2; maximum, 144; minimum, 2).

When the new interpretations were compared with the old reports, 977 scores of 1, 2, or 3 (insignificant discrepancies), and 23 scores of 4 or 5 (significant discrepancies) were assigned. During the final review process of all scores of 4 or 5, the third independent-expert neuroradiologist changed 1 score of 4 to a score of 1 and 2 scores of 4 to scores of 2. The resulting final distribution of scores along with their relative percentages is depicted in Table 2. In 87.6% (876/1000), there were no changes recommended to the report (score of 1).

Of the 20 cases with scores of 4 and 5, eight were CTs (7 of the brain, 1 CT angiogram) and 12 were MR images (9 of the brain, 1 of the neck, 1 of the spine, and 1 MR venogram). Of the 20 "misses," 7 cases were classified as vascular (4 aneurysms, 2 arterial occlusions, and 1 venous clot), 9 as neoplasms (5 dealing with the extent of disease progression and 4 new tumors), 2 as congenital (1 encephalocele, 1 gray matter heterotopia), and 2 as artifacts (Table 3).

Of the 100 (100/1000, or 10% of total) randomly selected reviewed cases, 8 (8/100 or 8%) were studies with normal find-

Table 3: Discrepant cases (scores of 4 or 5)

Discrepancy Category	Discrepancy Score (4 or 5)	Imaging Study Type
Vascular[a]	5	MR brain
Vascular[a]	4	CT angiogram
Vascular[a]	4	MR brain
Vascular[a]	4	MR brain
Vascular[a]	4	MR brain
Vascular[a]	4	MR brain
Vascular[a]	4	MR venogram
Congenital	5	CT brain
Congenital	5	CT brain
Neoplasm[b]	5	CT brain
Neoplasm[b]	5	MR brain
Neoplasm[b]	5	MR brain
Neoplasm[b]	4	CT brain
Neoplasm[b]	4	CT brain
Neoplasm[b]	4	MR brain
Neoplasm[b]	4	MR brain
Neoplasm[b]	4	MR neck
Neoplasm[b]	4	MR spine
Artifacts	5	CT brain
Artifacts	4	CT brain

[a] Aneurysms or venous clots.
[b] New lesions or the extent of disease progression.

Table 4: Individual discrepancy rates

Reviewer	Total Studies Read	Scores of 4 and 5	Discrepancy Rate
A	26	2	7.7%
B	16	1	6.3%
C	22	1	4.5%
D	45	2	4.4%
E	49	2	4.1%
F	59	2	3.4%
G	60	2	3.3%
H	111	3	2.7%
I	58	1	1.7%
J	68	1	1.5%
K	69	1	1.4%
L	114	1	0.9%
M	144	1	0.7%
N	4	0	0.0%
O	84	0	0.0%
P	65	0	0.0%
Q	2	0	0.0%
R	4	0	0.0%

ings (no abnormalities) and 92 (92/100 or 92%) had positive radiographic findings. There were 33 (33%) neoplastic, 23 (23%) vascular (stroke, aneurysms), 18 (18%) iatrogenic (tumor resection, ventricular shunt placement), 11 (11%) infectious/inflammatory (encephalitis, abscess, demyelinating disease), 7 (7%) traumatic, 3 (3%) congenital, 2 (2%) degenerative (spine), and 2 miscellaneous disease cases. Therefore, 2 of 30 (6.7%; 95% CI, 0%–15.8%) congenital abnormalities, 7 of 230 (3.0%; 95% CI, 0.8%–5.3%) vascular abnormalities, and 9 of 330 (2.7%; 95% CI, 0.9%–4.5%) neoplastic abnormalities were associated with a detection or interpretation discrepancy (no statistically significant difference, $P > .05$).

Two of the 20 cases (or 10%) with clinically significant discrepancies were originally read by a member of the faculty only, and the remaining 18 cases (or 90%) were read by a resident/fellow and faculty. Of the 100 randomly selected reviewed cases, 23 (23%) were originally reviewed by a staff neuroradiologist and 77 (77%) were read by a resident/fellow and staff. Thus the hypothetical clinically significant discrepancy rate for studies read by faculty only was 0.9% (2 in 230) and for studies read by faculty and resident/fellow was 2.3% (18 in 770).

Of the 1000 studies, 400 (40.0%) were CT and 586 (58.6%) were MR imaging examinations, and 14 (1.4%) did not have their imaging type noted by the reviewers. Of the CT studies, 284 (71.0%, 284/400) examined the brain/head and neck or were CT angiograms, 23 (5.8%, 23/400) looked at the spine, 73 (18.3%, 73/400) imaged bony facial or neck structures (sinuses, orbits, facial maxillary, temporal bone), and 20 (5.0%, 20/400) cases were labeled by the reviewers as CT image datasets without further specification. Of the MR imaging studies, 471 (80.4%, 471/586) imaged the brain/head and neck or were MR angiograms; 65 (11.1%, 65/400) examined the spine; 11 (1.9%, 11/586) looked at the orbits, neck, or face; and 39 (6.7%, 39/586) other MR imaging studies did not have the region of interest specified by the reviewers.

For all CT studies, the rate of clinically significant detection or interpretation discrepancies (scores of 4 or 5) was 2.0% (8 in 400; 95% CI, 0.6%–3.4%); for all MR imaging studies, it was 2.0% (12 in 586; 95% CI, 0.8%–3.2%). There were no statistically significant differences between CT and MR imaging discrepancy rates ($P = .88$). For all images of the brain/head and neck, the discrepancy rate was 2.5% (19 in 755; 95% CI, 1.4%–3.6%); for all images of the spine, it was 1.1% (1 in 88; 95% CI, 0.0%–3.3%) ($P = .48$). No discrepant cases were noted among the imaging studies of the bony facial structures, orbits, and face.

Overall, the neuroradiology service had a 2.0% (20 in 1000; 95% CI, 1.1%–2.9%) rate of clinically significant detection or interpretation discrepancies (scores of 4 or 5) when images were reread (Table 2). For the 18 neuroradiologists who dictated the original reports, the discrepancy rate ranged from 0% (minimum) to 7.7% (maximum) and the SD was 2.3%. The 3 highest discrepancy rates among the 18 neuroradiologists were 7.7% (2 in 26 studies), 6.3% (1 in 16 studies), and 4.5% (1 in 22 studies); the discrepancy rates for the top 3 readers who had the greatest number of studies reviewed were 0.7% (1 in 144 studies), 0.9% (1 in 114 studies), and 2.7% (3 in 111 studies). Table 4 demonstrates individual discrepancy rates.

The 18 neuroradiologists in this study had an average of 9.9 ± 8.7 years of experience in the field (range, 1–28 years). There was no relationship between the years of experience and discrepancy rates when the neuroradiologists with scores of 4 or 5 were compared with those without scores of 4 or 5 ($P = .11$). The Spearman rank correlation coefficient for years of experience and discrepancy rate was 0.25 with a P value of .30 (not significant). Because of the low numbers of clinically significant discrepancies for each neuroradiologist, these calculations were underpowered, contributing to the lack of statistically significant differences.

Twelve of 18 neuroradiologists who filed the original reports completed their neuroradiology fellowships at our insti-

tution, and the remaining 6 completed their training at other fellowship programs. The discrepancy rate was 2.2% (95% CI, 1.0%–3.4%) and 2.8% (95% CI, 0.4%–5.1%) for internally trained and outside-trained staff, respectively ($P = .64$).

There was no relationship between the number of studies read and the discrepancy rate when the neuroradiologists with scores of 4 or 5 were compared with those without scores of 4 or 5. Reviewers without a single discrepant case read on average 31.8 ± 39.6 studies (range, 2–84), and reviewers with at least 1 discrepant case read, on average, 64.7 ± 38 studies (range, 16–144) ($P = .12$).

Discussion

In this study, 1000 neuroradiology examinations read by fellowship-trained neuroradiologists on staff at a major academic medical center were subjected to peer review. We found a 2.0% (95% CI, 1.1%–2.9%) rate of clinically significant detection or interpretation discrepancy between the original report and the second opinion review for the whole neuroradiology division. The clinically significant discrepancy rate ranged from 0% to 7.7% for individual neuroradiologists. There was no statistically significant difference between error rates for CT and MR imaging studies and between error rates for brain/head and neck and spine imaging. Not a single clinically significant discrepancy (score of 4 or 5) was recorded for CT and MR imaging studies of the bony facial structures, orbits, and face. Also, we found no relationship between the years of experience in the field or the number of total cases read and missing a case.

Our overall error rate was somewhat lower but comparable with the 5% rate predicted by Garland and with the 3%–6% discrepancy rates observed by other researchers for non-neuroradiologic studies.[3-5,12-14] Our divisional discrepancy rate was also lower than the recently reported 4.2% major disagreement rate seen between staff neuroradiologists and residents interpreting emergent neuroradiology MR imaging examinations.[16] Our relatively low discrepancy rate was observed despite factors that, one would expect, should have increased the radiologic error rate. These factors were inclusion of complex imaging studies (cross-sectional imaging only; MR imaging > CT), a sample case mix with high prevalence of disease, and hindsight bias—knowledge of how the case evolved with time—resulting from the availability of the follow-up examination. Higher discrepancy rates for CT and MR imaging compared with plain film and for body regions with complicated anatomy have been documented.[3,17] Also, as the disease prevalence increases, so does the expected error rate, approaching Garland's rate of 30% for case mixes with 100% disease prevalence.[4-6,12] Borgstede et al[3] showed that academic medical centers compared with community hospitals had error rates approximately 1.5% higher, presumably due to greater disease prevalence. In this study, the overall disease prevalence was 92%.

There were also specific factors that may have lowered our discrepancy rate. Because we only reviewed cases for which follow-up studies were performed, we not only increased the number of positive findings (requiring follow-up) but also likely selected certain types of imaging examinations and disease. Wong et al[17] found an error rate of 1.09% (individual radiologists ranged from 0.7% to 1.41%) among 10 teleradiologists who read "off-hours" emergent studies.[17] In that study, 10 types of examinations composed nearly 90% of all cases reviewed, which led the authors to conclude that the low error rate may have resulted from the teleradiologists being well-attuned to the specific findings and abnormalities of the limited kinds of examinations.[17]

Our facility is a large tertiary care academic medical center, and we reviewed examinations read by 18 fellowship-trained neuroradiologists. Borgstede et al[3] showed a decreasing discrepancy rate for large compared with small institutions and observed a reduction of the discrepancy rate of 0.5% for each 10 additional radiologists on staff. Finally, similar to other quality-improvement projects that are based on peer review, our study may have been affected by under-reporting bias resulting from a tendency of fully trained professionals not to disclose the errors of their colleagues, especially in an environment that measures and compares individual performance. Because our secondary case review was fully incorporated into the daily clinical work flow, the names of the original study reviewers were not masked in any way. This may have contributed to under-reporting or misclassification of the discrepancies of the original report (significant versus nonsignificant).

The medical-legal literature divides missed radiologic findings into errors due to negligence and errors not due to negligence.[2] Only the radiologic errors that result from "a breach of the standard of medical care" are deemed negligent.[2] A thorough review of medical-legal cases by Berlin[2] suggests that courts do permit certain mistakes, that a radiologist cannot always be perfect, and that even the most careful and scrupulous examination of a radiograph by 2 different radiologists or by the same radiologists at different time points may result in a clinically significant mistake. In 1959, Garland[5] attributed such "inevitable" errors to the mysterious "human equation." Since then, researchers have looked at many relevant factors and their impact on reading accuracy, including the effect of the reading-room environment, multitasking distractions, the availability of clinical history, the availability of previous reports and studies, the duration of search time, and other factors.[6,18-20]

In this study, we did not measure or control for such factors. All reviewed dictations were performed with commercially available software and hardware, and staff were individually responsible for proofreading their own reports. At our institution, on average, faculty reads 25% of cases on their own and 75% of cases with a resident or a fellow. In our study, the hypothetic rate of clinically significant discrepancies for studies read by resident/fellow and faculty was higher than that for studies read by faculty only. There is also substantial evidence in favor of the phenomenon of satisfaction of search, whereby a reader stops looking for additional abnormalities once a certain number of findings have been reached.[21,22]

Training and experience are essential for the development and maintenance of an accurate image interpreter. Eng et al[23] compared the interpretations of plain radiographs with known findings between radiology and emergency medicine physicians on staff and radiology and emergency medicine residents. They found that radiologists on staff were better than radiology residents, radiology residents were more accurate than emergency medicine staff, and emergency medicine attendings erred less than emergency medicine residents. Sim-

ilarly, in subspecialty imaging, radiologists with fellowship training and focused experience outperformed board-certified general radiologists. Neuroradiologists, oncologic radiologists, and mammography-trained specialists were better at interpreting images within their areas of expertise than community radiologists with many years of experience.[24-26]

The converse argument also seems to be true. A study by Branstetter et al[27] documents that subspecialty radiologists outside of their field of expertise do worse than senior radiology residents when interpreting basic body films, further highlighting the importance of ongoing experience. Our results suggest that the process of learning through experience continues among neuroradiologists on staff and that even those who are considered "experts" by some standards[28] can still err. On the basis of the published literature, one may expect that fellowship-trained neuroradiologists with a greater number of years of clinical experience would make fewer mistakes compared with their "younger" colleagues. However, in our limited sample, we did not detect any statistically significant relationship between the clinically significant discrepancy rate and the number of years in practice beyond neuroradiology fellowship training.

Quality improvement projects throughout the country aim to better patient care by improving outcomes and/or providing more tangible benefits ("value") per dollar spent on care.[29,30] Such projects are typically associated with measuring either outcomes (eg, the risk-adjusted inpatient mortality rate post coronary artery bypass graft surgery) or clinical processes (eg, the percentage of patients with suspected pneumonia who receive antibiotics within 4 hours of presenting to the emergency department). These measurements can be used for internal continuous quality improvement initiatives like Six Sigma, Kaizen, or Toyota Production System or for external reporting and accountability, which are often associated with threshold-based rewards or sanctions.[29,31] Our study was a part of a quality assurance initiative in our department, which aimed to survey the rates of detection or interpretation discrepancies between original reports and second-opinion reviews among staff neuroradiologists. We did not focus on measuring clinical outcomes but instead assessed the process of radiographic interpretation with hopes that improvements in the interpretation consistency will have direct beneficial effects on the patient care outcomes. In our approach, to create a collaborative environment, we did not focus on rewarding or penalizing our staff. This practice was to shield the staff from any anxiety-provoking performance expectations and to shape an atmosphere in which sharing of experiences and knowledge is fostered and where a mistake is perceived as an educational opportunity.

Another motivator for our departmental quality-improvement initiative and this study is the recently added requirement for the maintenance of certification by the American Board of Medical Specialties of PQI. Our study is an example of initial data collection as one of the first steps of a PQI project.

The next step in the PQI process would be to obtain absolute accuracy rates (not just discrepancy rates) for the cases reviewed and to intervene with an initiative for improvement. Certainly, the wide range of values between reviewers (from 0% to 7.7%) suggests a closer look at the individuals with a higher rate of discrepant reports and the value of adding more cases to the review to increase power.

At our institution, we update and review faculty discrepancy rates semiannually and compare the individual rates with the department-wide rate. Staff neuroradiologists who are 2 SDs above the mean are required to take and pass an ABR-approved self-assessment module addressing the topics with the greatest number of errors. If on the subsequent quality assessment review, the faculty member continues to struggle in the same area, he or she is asked to complete a relevant course offered by a neuroradiology society, such as the American Society of Spine Radiology, the American Society of Pediatric Neuroradiology, the American Society of Functional Neuroradiology, the American Society of Interventional and Therapeutic Neuroradiology, and the American Society of Head and Neck Radiology.

This study has a number of limitations. Our 5-point scoring system, though validated in prior studies,[15] was somewhat arbitrary and may have failed to capture or may have misclassified some of the differences between the original reports and the second-opinion reviews. Also, we did not test the intra- and interobserver agreement of the scoring system, and any variability among the reviewers in assigning a clinically significant versus nonsignificant category to the perceived discrepancies could have affected our individual and overall clinically significant discrepancy rates. Nonetheless, this scale has been reported to show a modest 4.6% disagreement rate in a prior publication.[15] We had a low compliance rate. The 1000 neuroradiology studies examined herein are only a fraction of studies we expected to capture, given that we asked all of our neuroradiologists on staff to review 2 cases on each of their clinical days during a 17-month period.

Our study design likely introduced disease-selection bias and hindsight bias. We only included cases for which follow-up studies were being performed (ie, examinations without a prior comparison study were not re-assessed); this decision probably altered the distribution of diseases seen and types of examinations reviewed (eg, 586/1000 studies were MR images and 330/1000 studies had neoplastic pathologies). Hindsight bias may have resulted from the reviewers knowing how the disease process declared itself on the follow-up imaging and thus from falsely assuming that they would have perceived subtle changes on an earlier study.[32] However, both of these limitations, if anything, should have exaggerated the observed clinically significant discrepancy rate, yet the rate was appropriate on the basis of literature. Also, our goal was to devise a practical and efficient quality assurance program that enhances the daily workflow by encouraging faculty to review prior reports. Our design did not call for reading cases that faculty would not have otherwise read.

Another limitation has to do with the homogeneity of training of our staff neuroradiologists. Of the current 11 neuroradiologists on staff, 7 completed their fellowship training at our institution; of the 18 neuroradiologists who filed the original reports, 12 completed their neuroradiology fellowship at our institution. Thus, their imaging interpretation tendencies are likely similar and prone to reflect the same biases. However, in our sample of 1000 cases, we did not see a statistically significant difference between the clinically significant discrepancy rate among the internally trained and outside-

trained neuroradiologists. Additionally, as a part of the scoring-system validation and data normalization, our methodology called for 1 expert reviewer to take a final look at all (except for his own) the clinically significant discrepancies (scores of 4 and 5) and to make a final determination of the scores. Although the expert reviewer has 22 years of experience in the field and had previously worked with the scoring system used herein, this step may have introduced bias.

We also did not focus on follow-up of the discrepant readings to determine the "true diagnosis" because this study was focused on discrepancy rates, not accuracy. Thus, if both reviewers of the same study were incorrect in their interpretations, the agreement would be scored as no change, but the study would have been interpreted wrongly. In this study, we studied consistency; however, a follow-up study to assess accuracy rates, not just discrepancy rates, is warranted.

This study looked at the first consecutive 1000 neuroradiology studies that had undergone a secondary peer review as a part of a quality improvement initiative in our division. Due to the low overall frequency of the clinically significant discrepancies in our sample, for some of the subanalyses, such as the association between the years of experience and clinically significant discrepancies and the number of studies read and clinically significant discrepancies, the statistical calculations were likely underpowered.

Most important, because our institution is a major academic medical center with a large tertiary care patient population base, the percentage of studies with normal findings is <10%. The acuity of the cases and their complexity (with multiple diseases and/or unusual diseases), given the referral base, may have influenced our results. In a similar fashion, the advanced technologies used (diffusion-weighted imaging, diffusion tensor imaging, and MR spectroscopic imaging) on the latest high-quality scanners, paired with advanced sequences and reconstructions, may not reflect the study types performed in other radiology practice settings. In a practice in which more neuroradiology findings are normal for typical indications such as headaches or lower back pain, the discrepancy rates may be less. In other words, there is a selection bias for cases with positive findings in our sample.

Conclusions

We found a 2.0% rate of clinically significant detection or interpretation discrepancy between original dictations and second-opinion reviews of neuroradiology cases among fellowship-trained neuroradiologists at a university hospital. Our results may have limited generalizability because we only reviewed cases for which follow-up studies were being done; thus, we potentially introduced disease-selection bias. Also, our study design may have introduced hindsight bias by allowing the reviewers to see how the cases evolved with time. Our study is an example of 1 step in the PQI process that could serve as a blueprint for collection of data before an intervention to improve the homogeneity of interpretation quality. Accuracy rates must also be addressed for the best patient care.

References

1. Turner DA. **Observer variability: what to do until perfect diagnostic tests are invented.** *J Nucl Med* 1978;19:435–37
2. Berlin L. **Radiologic errors and malpractice: a blurry distinction.** *AJR Am J Roentgenol* 2007;189:517–22
3. 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
4. Garland LH. **On the scientific evaluation of diagnostic procedures.** *Radiology* 1949;52:309–28
5. Garland LH. **Studies on the accuracy of diagnostic procedures.** *Am J Roentgenol Radium Ther Nucl Med* 1959;82:25–38
6. Berlin L. **Accuracy of diagnostic procedures: has it improved over the past five decades?** *AJR Am J Roentgenol* 2007;188:1173–78
7. Herman PG, Gerson DE, Hessel SJ, et al. **Disagreements in chest roentgen interpretation.** *Chest* 1975;68:278–82
8. Austin JH, Romney BM, Goldsmith LS. **Missed bronchogenic carcinoma: radiographic findings in 27 patients with a potentially resectable lesion evident in retrospect.** *Radiology* 1992;182:115–22
9. Janjua KJ, Sugrue M, Deane SA. **Prospective evaluation of early missed injuries and the role of tertiary trauma survey.** *J Trauma* 1998;44:1000–06, discussion 1006–07
10. Anderson RE, Hill RB, Key CR. **The sensitivity and specificity of clinical diagnostics during five decades: toward an understanding of necessary fallibility.** *JAMA* 1989;261:1610–17
11. Roosen J, Frans E, Wilmer A, et al. **Comparison of premortem clinical diagnoses in critically ill patients and subsequent autopsy findings.** *Mayo Clin Proc* 2000;75:562–67
12. 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
13. 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
14. 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
15. Zan E, Yousem DM, Carone M, et al. **Second-opinion consultations in neuroradiology.** *Radiology* 2010;255:135–41
16. Filippi CG, Schneider B, Burbank HN, et al. **Discrepancy rates of radiology resident interpretations of on-call neuroradiology MR imaging studies.** *Radiology* 2008;249:972–79
17. Wong WS, Roubal I, Jackson DB, et al. **Outsourced teleradiology imaging services: an analysis of discordant interpretation in 124,870 cases.** *J Am Coll Radiol* 2005;2:478–84
18. Oestmann JW, Greene R, Kushner DC, et al. **Lung lesions: correlation between viewing time and detection.** *Radiology* 1988;166:451–53
19. Berlin L. **Comparing new radiographs with those obtained previously.** *AJR Am J Roentgenol* 1999;172:3–6
20. Loy CT, Irwig L. **Accuracy of diagnostic tests read with and without clinical information: a systematic review.** *JAMA* 2004;292:1602–09
21. Samuel S, Kundel HL, Nodine CF, et al. **Mechanism of satisfaction of search: eye position recordings in the reading of chest radiographs.** *Radiology* 1995;194:895–902
22. Berbaum KS, Franken EA Jr, Dorfman DD, et al. **Role of faulty visual search in the satisfaction of search effect in chest radiography.** *Acad Radiol* 1998;5:9–19
23. Eng J, Mysko WK, Weller GE, et al. **Interpretation of emergency department radiographs: a comparison of emergency medicine physicians with radiologists, residents with faculty, and film with digital display.** *AJR Am J Roentgenol* 2000;175:1233–38
24. Erly WK, Ashdown BC, Lucio RW 2nd, et al. **Evaluation of emergency CT scans of the head: is there a community standard?** *AJR Am J Roentgenol* 2003;180:1727–30
25. Dudley RA, Hricak H, Scheidler J, et al. **Shared patient analysis: a method to assess the clinical benefits of patient referrals.** *Med Care* 2001;39:1182–87
26. Sickles EA, Wolverton DE, Dee KE. **Performance parameters for screening and diagnostic mammography: specialist and general radiologists.** *Radiology* 2002;224:861–69
27. Branstetter BF 4th, Morgan MB, Nesbit CE, et al. **Preliminary reports in the emergency department: is a subspecialist radiologist more accurate than a radiology resident?** *Acad Radiol* 2007;14:201–06
28. McCarron MO, Sands C, McCarron P. **Quality assurance of neuroradiology in a district general hospital.** *QJM* 2006;99:171–75
29. Lilford RJ, Brown CA, Nicholl J. **Use of process measures to monitor the quality of clinical practice.** *BMJ* 2007;335:648–50
30. Hayward RA. **Performance measurement in search of a path.** *N Engl J Med* 2007;356:951–53
31. Lindenauer PK, Remus D, Roman S, et al. **Public reporting and pay for performance in hospital quality improvement.** *N Engl J Med* 2007;356:486–96
32. Harvey JA, Fajardo LL, Innis CA. **Previous mammograms in patients with impalpable breast carcinoma: retrospective vs blinded interpretation—1993 ARRS President's Award.** *AJR Am J Roentgenol* 1993;161:1167–72