Rates of Disagreement in Imaging Interpretation in a Group of Community Hospitals[1]

Robert L. Siegle, MD, Evelyn M. Baram, PhG, JD, Stewart R. Reuter, MD, JD
Ewell A. Clarke, MD, Jack L. Lancaster, PhD, C. Alex McMahan, PhD

Rationale and Objectives. Prospective studies of radiologists' interpretations of selected radiographs reported 20–40 years ago indicated error rates of 30% and higher. The authors retrospectively evaluated the interpretations of groups of radiologists and determined a range of rates of disagreement in interpretation. Quality assessment or recredentialing may add to the importance of such studies in the future.

Materials and Methods. Over a 7-year period, a team of radiologists reviewed imaging interpretations in the radiology departments of six community hospitals. Each review, which lasted about 3 days, included evaluation of the interpretations of a 3%–4% sample of the images read by the radiologists at these hospitals. Reading errors were quantitated and evaluated qualitatively.

Results. In a review of over 11,000 images read by 35 radiologists, the authors found a 4.4% mean rate of interpretation disagreement; only one radiologist had a mean rate above 8%. Qualitative analysis of the interpretation errors revealed a mean rate of 3.0% of errors that were considered to be below an acceptable standard of care. Radiologists whose errors included a relatively high proportion of false-positive findings tended to make relatively fewer total errors.

Conclusion. Rates of disagreement for a broad range of studies that radiologists interpret in a community hospital setting appear to be far lower than earlier studies on selective radiographs indicated.

Key Words. Diagnostic radiology, observer performance; images, interpretation; quality assurance.

Radiologists make a certain inescapable minimum number of interpretative errors, no matter how diligent they may be (1,2). Reports dating as far back as World War II yielded error rates of 30% and higher. These studies, however, did not mimic everyday practice but rather were based largely on radiographs that were selected for the presence of disease and then given to groups for interpretation. We retrospectively evaluated the interpretations of groups of radiologists in a community hospital setting and determined a range of rates of disagreement in interpretation. Demands for improved quality assessment and consideration for recredentialing make this work relevant to the practice of radiology today.

MATERIALS AND METHODS

Over a 7-year period, a team of radiologists reviewed radiology departments in six community hospitals. As part of the review a sequential sample of the radiologists' studies over the course of the previous year were interpreted by the team and the findings compared with those of the original readings. The studies were stratified by type into the following categories: plain radiography, fluoroscopy, angiography, ultrasound (US), computed tomography (CT), mammography, nuclear medicine, and magnetic resonance (MR) imaging. The relative number of types of studies reviewed reflected the distribution of studies obtained in the

Acad Radiol 1998; 5:148-154

[1] From the Departments of Radiology (R.L.S., S.R.R., E.A.C., J.L.L.) and Pathology (C.A.M.), University of Texas Health Science Center, 7703 Floyd Curl Dr, San Antonio, TX 78284-7800; and the American Medico-Legal Foundation, Philadelphia, Pa (E.M.B.). Received July 21, 1997; accepted and revision requested August 25; revision received September 26. **Address reprint requests to** R.L.S.

Table 1
Classification of Errors of Interpretation

Class	Description
I	The finding or diagnosis could not be expected to be identified by a general radiologist.
II	The finding or diagnosis was difficult to see.
III	The finding or diagnosis should have been observed by most radiologists.
IV	The finding or diagnosis should have been observed by any physician.
V	Overreading: The finding or diagnosis was not substantiated by the team's review.

Table 2
Group Rates of Disagreement

Hospital	Disagreement Rate (%)	Substandard Rate (%)*
A	5.2	3.3
B	5.4	3.4
C	3.0	NA
D	5.1	3.1
E	2.9	1.8
F	4.8	NA

*NA= not analyzed.

department during that year. The team, which included as many as eight radiologists, had the same nucleus of three radiologists (R.L.S., S.R.R., and E.A.C.) for all of the hospital reviews; additional radiologists were added depending on hospital size and the number of images to be reviewed. After the first review, a target was set to evaluate a 3% representative sample of each radiologist's work. The target was set to achieve a degree of reproducibility within the constraints of physician time and hospital finances. As many as 3,400 images were evaluated in a review. Some of the radiologists had fewer studies reviewed than others. This does not mean that they interpreted fewer studies in the course of a year, but rather that some of the groups had private offices, and studies obtained in these offices were not part of the hospital reviews.

The reviewing radiologists interpreted plain radiographs and those subspecialty studies in their own areas of expertise. The reviewers made their interpretations without the benefit of the patient history or the original report. If the reviewer concurred with the findings of the original report, no further action was taken. If the reviewer disagreed with the report, the team of reviewers jointly interpreted the study. The team's consensus was accepted as correct. For the last four hospital reviews the team also attempted to make a qualitative evaluation of the errors of interpretation and then classify the errors by severity (Table 1).

These classes of severity are slightly modified and expanded from those in a classification system that is used by the American College of Radiology (3,4). The team considered class III and IV errors to be below an acceptable standard of care. Disagreement rates were determined for each radiologist.

Each of these departmental reviews occurred at the request of hospital administration and was performed with the cooperation of the hospital radiologists and the remainder of the professional staff. The departmental reviews were usually requested when the administrators perceived that problems existed either among the radiologists or between the radiologists and other specialists, and when the administration and professional staff had been unable to resolve the issues intramurally. The initial request for assistance was made though the American Medico-Legal Foundation, a nonprofit, peer-review organization that reviews hospital departments.

In addition to the review of radiologic interpretations, the departmental reviews included interviews with the hospital administrators, radiologists, and other physicians who practiced at the hospital. The team also evaluated the quality assurance plans, radiology equipment, film and procedure techniques, and alleged instances of inadequate care. The review offered solutions for political and administrative problems and suggestions for equipment and physical plant changes along with analyses of quality of care issues. The team provided a preliminary summary in the form of an exit interview with all interested parties, and each completed review was presented subsequently in written form to the hospital administration. The scope of this article is limited to evaluation of rates of disagreement in radiologic interpretation.

RESULTS

We evaluated 11,094 interpretations made by 35 radiologists in the six departments studied. The overall rates of disagreement for the six radiology groups ranged from

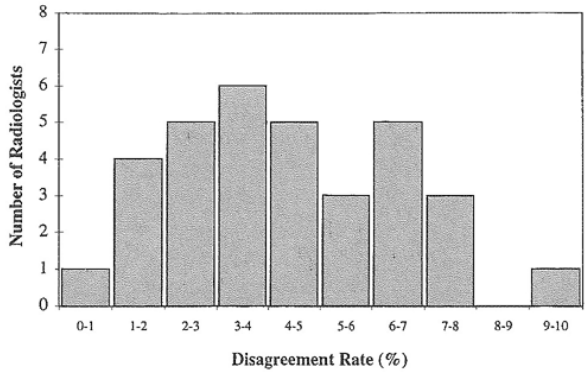

Figure 1. Histogram shows the number of radiologists at each range of percentage disagreement.

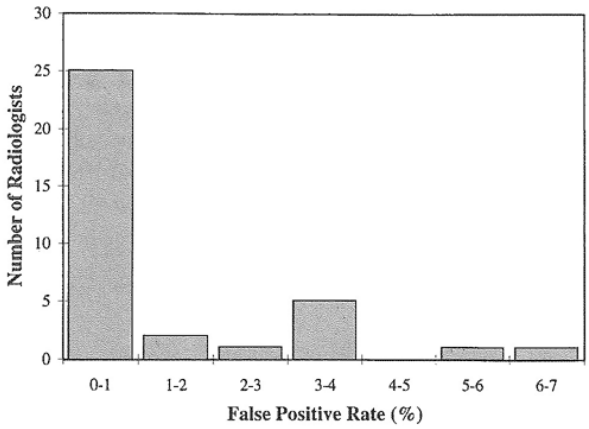

Figure 2. Histogram shows the number of radiologists at each range of percentage false-positive readings.

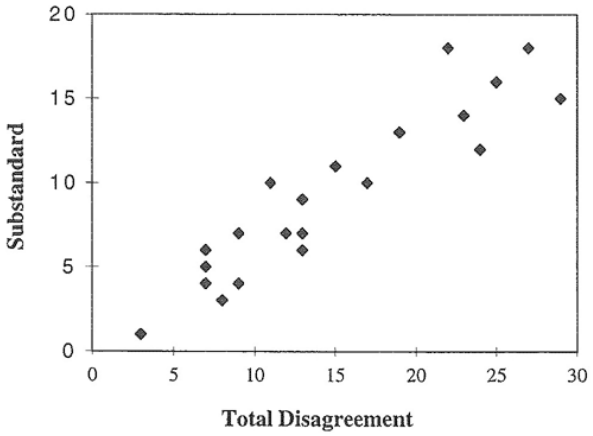

Figure 3. Graphic comparison of the number of errors made by each radiologist with the number of substandard errors. R^2 = 0.856, slope = 0.609, intercept = 0.261.

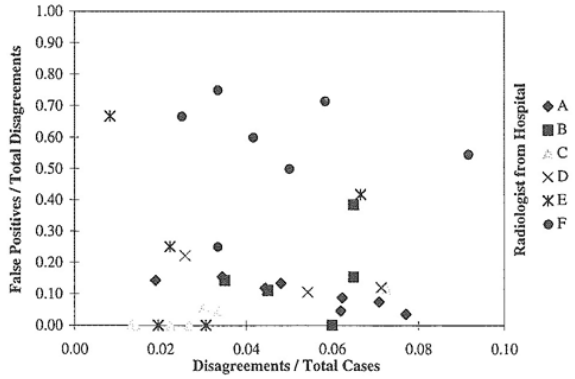

Figure 4. Graphic comparison of the fraction of false-positive readings to total number of disagreements plotted against the fraction of disagreements to total number of cases.

2.9% to 5.4%. The rates for substandard interpretations ranged from 1.8% to 3.4% (Table 2).

The breakdown for disagreement rates for individual radiologists is given in Table 3. Because the radiologists interpreted different numbers of images, an unweighted mean of the radiologists was computed along with a one-sided 95% confidence interval. Because the cases in which the reviewer agreed with the original evaluation were not separated by type of finding (positive or negative), we could not compute the κ statistic for agreement. The individual rates of disagreement ranged from 0.8% to 9.2%. A histogram showed an approximation of a normal distribution and a mean rate of disagreement of 4.4% (Fig 1). Only one radiologist had a rate above 8.0%. Substandard interpretation rates ranged from 0.3% to 5.1%, with a mean of 3.0%. Disagreement on the side of overreading ranged from 0% to 6.3%, with a mean of 1.2%; only two radiologists had a rate above 4% (Fig 2).

The relationship between disagreements that were considered to reflect substandard care and total number of disagreements is plotted in Figure 3. Figure 4 provides a means of analyzing disagreements on the side of overreading (false-positive results). The fraction of false-positive readings to total number of disagreements is plotted against the fraction of total disagreements to all cases. The radiologists are identified according to hospital. The relation of the fraction of false-positive readings to total number of disagreements, as well as the differences between hospitals, was evaluated by using analysis of covariance. Multiple comparisons were carried out with the Scheffe method, and the fraction of false-positive read-

Table 3
Disagreement Rates for Individual Radiologists

Radiologist	Total No. of Studies Reviewed	No. False-Negative	No. False-Positive	No. Substandard	Disagreement Rate (%)	Upper Confidence Limit (%) *
A-1	369	21	2	14	6.23	8.71
A-2	383	15	2	10	4.44	6.58
A-3	381	25	2	18	7.09	9.64
A-4	355	21	1	18	6.20	8.73
A-5	312	13	2	11	4.81	7.31
A-6	371	6	1	4	1.89	3.51
A-7	376	28	1	15	7.71	10.37
A-8	378	11	2	7	3.44	5.41
B-1	200	11	2	9	6.50	10.14
B-2	200	6	1	5	3.50	6.47
B-3	200	8	5	6	6.50	10.14
B-4	200	12	0	7	6.00	9.54
B-5	200	8	1	7	4.50	7.72
C-1	372	24	3	NA	7.26	9.87
C-2	624	20	1	NA	3.37	4.81
C-3	142	2	0	NA	1.41	4.37
C-4	603	8	0	NA	1.33	2.38
C-5	601	16	0	NA	2.66	4.02
C-6	596	17	1	NA	3.02	4.45
C-7	541	12	0	NA	2.22	3.57
D-1	350	7	2	4	2.57	4.44
D-2	350	22	3	16	7.14	9.83
D-3	350	17	2	13	5.43	7.86
E-1	360	7	0	6	1.94	3.62
E-2	360	14	10	12	6.67	9.25
E-3	360	6	2	3	2.22	3.97
E-4	360	11	0	10	3.06	5.01
E-5	360	1	2	1	0.83	2.14
F-1	120	5	6	NA	9.17	14.72
F-2	120	2	5	NA	5.83	10.68
F-3	120	2	3	NA	4.17	8.56
F-4	120	3	1	NA	3.33	7.47
F-5	120	1	2	NA	2.50	6.33
F-6	120	1	3	NA	3.33	7.47
F-7	120	3	3	NA	5.00	9.63

*NC = not calculated
†One-sided 95% confidence interval.

ings to total disagreements was transformed by using angular transformation before statistical analysis. A weighted analysis was used to account for the different number of images reviewed for different radiologists. The ratio of false-positive readings was unrelated to the disagreement rate ($P = .27$). The ratio for hospital F as a group, however, was significantly different from the ratios for the other hospitals ($P < .05$).

DISCUSSION

The most frequently cited works in this field were prepared several decades ago. The first authority in the field was L. Henry Garland, who wrote four articles between 1949 and 1960. The first extensive review by Garland in 1959 (6) included references to earlier limited studies of error rates by radiologists, error rates in medical fields other than radiology, displays of optical illusions, and relevant biblical verses. He cited his own study of 8,931 sanatorium radiographs (presumably all positive) for agreement on improvement, worsening, or no change and found interobserver disagreement at 30% and intraobserver disagreement at 21% (6). Similarly, his analysis of 14,867 photofluorograms disclosed a 39% rate of underreading and a 1.2% rate of overreading (6).

Herman and Hessel (7) obtained 100 positive hospital chest radiographs and developed an idealized report for each based on initial interpretation, review of clinical information, synthesis of test reports, and a final panel judgment. The report of each tested radiologist was analyzed to see if it had all the points made in the idealized report. They found that 26% had statements that included important or potentially important errors.

Lehr et al (8) reported error rates of 30.9% and 33.4% when 16 radiologists were each shown 2,145 test radiographs and their 35-mm equivalent reproductions. The study included 18% normal radiographs. For these alone the error rate was 19%.

Rhea et al (9) described 161 initial errors in radiographic interpretations performed in the emergency department by residents. The errors were identified the following day during staff review. Twenty-five percent of all the studies had important abnormalities related to the clinical question (affecting clinical care), and 7% had unrelated important abnormalities. "The previously reported error rate in an observer's interpretation of an image is about 30% (20%–40%) and is confirmed by this study" (9). In fact, these error rates should be more clearly defined as false-negative rates, since these studies (aside from a small fraction of the cases of Lehr et al) all represent interpretation of images with known positive findings (6–9). Garland himself reflected on this difference in a small-print footnote in the report of his own study (6) and indicated that a 5% error rate should be expected if surveys evaluate the whole spectrum of radiographs that a radiologist interprets.

The meaning of the term "error rate" is not intuitively obvious. The above cited papers used the statistical definition of error rate and calculated it by dividing the number of false-negative interpretations by the number of reference standard positive cases and multiplying by 100. Thus, if there were three positive cases in a series of 100, and the radiologist interpreted two of these cases correctly as positive and one incorrectly as negative and all 97 negative cases correctly as negative, the error rate would be 33% even though 99 of the 100 studies were correctly interpreted. Some radiologists, not to mention clinical colleagues and the lay public, hear of an error rate in radiology of "30% or greater" and immediately assume that we misinterpret a third of our examinations.

Renfrew et al (10) recognized this distinction in terminology and referred only to false-negative readings in their review drawn from problem case conferences. They noted that "our findings agree with those of prior studies that noted failure to detect lesions in 25%–32% of cases." Yet, another important review of the same material failed to make this distinction (11). "Summarized in depth previously by Berlin, as well as others, all have confirmed Garland's original data that error rates among radiologists hover in the 30% range" (11).

The rates of disagreement in our study confirm the impressions of Garland. In the six broad surveys of radiology departments, the mean rate of disagreement between our review group and the radiologists being evaluated was 4.4%, which was close to Garland's expected 5%. One of the key reasons for each of these radiology department evaluations, however, was to evaluate allegations of inadequacies of interpretation. The consensus was that only one radiologist of the 35 evaluated had a rate that was above an acceptable level. Such a determination could be made only subjectively, since no standards for surveys exist and our own evaluations are necessarily subjective.

More specific analyses, such as receiver operating characteristic curves or any other statistic that would place these results in the context of other published studies, were not feasible, since true-positive results and true-negative results were lumped together, and neither a true-

positive nor a true-negative fraction was established during our studies (12). Determining such fractions would require reaching consensus on all true-positive and true-negative results; in essence, that requires consensus on all image interpretations, not just those that are false-positive or false-negative. Such an evaluation of even 3% of a practice would be prohibitive in terms of time and money. The absence of the true-negative fraction also means that we did not calculate error rates in the statistical sense for the radiologists that we reviewed but rather rates of underread and overread studies summated to rates of disagreement with the reviewer group. Several additional correlations, however, can be made from the above data. Figure 3 demonstrates that a fairly predictable ratio exists between the number of reading disagreements and the number that were considered substandard. We did not, however, attempt to draw any conclusions about the number of a radiologist's errors that were considered substandard. The assignment of reading errors to classes considered substandard (classes III and IV) was necessarily subjective and in a sense arbitrary in imposing a categorization scheme on a continuous variable.

The plot in Figure 4, which shows a statistically significant difference in overreading for one hospital, raises an interesting question. It suggests that the group-reading milieu may have a marked effect on individual radiologists and cause them to alter interpretation patterns that had been developed in diverse training programs. We could find no unique biases or local environmental factors that would account for the increased overreading in hospital F.

Several changes occurred in the format of our evaluation over the 7-year period. The qualitative evaluation of errors developed after several interpretation errors were identified in the first two studies that merited the immediate attention of the hospital's professional staff with the intent of changing a specific patient's care. During the last departmental review we thought that two radiologists made repetitive errors in a specific field. This led us to request an additional number of studies in this field for the two radiologists that were outside the scope of the original survey. The additional studies confirmed the original impressions and led to one-on-one conferences with these radiologists to discuss the systematic errors in interpretation. These additional studies are not included in the tabulated error count in Table 3.

Several weaknesses exist in this type of evaluation of radiologic practice. First, the reviewers were not all the same during the course of the six reviews; therefore, the sensitivity of the reviewers was not constant although the consensus review of all purported interpretive errors probably compensated for interpersonal variations in sensitivity. Second, since no "gold" standard existed, the intrinsic false-negative rate for each reviewer also affects the statistics. This means that the rates for underreading for the radiologists could be slightly higher than those reported above. Third, radiologists do not all read the same proportion of the studies in a group. Therefore, reviewers may have evaluated a larger than proportional share of one radiologist's work in a group, or perhaps more of one type of study than is representative of that radiologist's yearly work.

Different examination types probably show different degrees of disease prevalence. For example, there is probably a higher incidence of disease depicted on CT scans than on mammograms, and so there is the potential for more false-negative results for a radiologist who reads a disproportionate share of CT scans relative to mammograms in that hospital. In a broader sense, radiologists in an office-based practice will probably see fewer studies with positive findings than those in a hospital-based practice. In hospitals where a radiologist performed very few of a given type of study, we were informed of this and attempted to compensate by reviewing fewer of those studies for the given radiologist. In the six community hospital practices studied, we noted that most of the work was shared evenly.

Strengths in this process are also evident. First, radiologist "outliers" can be identified in a reasonably rapid manner. Second, evaluation by subspecialty-trained reviewers identifies areas of weakness and provides direction for improvement, which ultimately improves the quality of care. Third, the in-house review provides assistance to the administrative and medical staff in solving political and administrative problems in their radiology departments.

The rates of disagreement established for community hospitals are probably not transferable to teaching hospitals. Teaching hospital statistics are affected by a greater number of patients with positive findings, which results in a greater proportion of positive images. As the positive rate approaches 100%, the error rate would approach Garland's 30%, but these numbers are tempered by the fact that many teaching hospital studies today are both double read and interpreted by radiologic subspecialists far more commonly than in Garland's time.

IMPLICATIONS FOR QUALITY ASSESSMENT

A study of this nature is feasible only if the examinees agree to it. No coercive agency exists at this time to compel such a general review, although third-party payers

have been a force behind some reviews (13). Alternatively, error rates can be determined in much the same way today as they were in Garland's early work. This involves the use of a prepackaged set of positive studies, which are much the same as the oral radiology board examinations. The clear drawback of this method is that such a selection of images does not reflect the practice milieu of the radiologists being evaluated.

Quality assessment of actual practice situations already exists in mammography. This, however, is focused on an area in which outcomes are easier to track than in most of radiology, and it is mandated by the Mammography Quality Standards Act of 1992 (14). Quality assessment in other areas is more difficult if it is to be more comprehensive than our review but remain within reasonable time and financial constraints. A magnetic resonance imaging quality assessment study of samples from 33 centers that involved 369 procedures was performed by three panels of three radiologists each (15). They accumulated much useful data, but the study took more than 2 years to perform.

If quality assessment is to be part of recredentialing in the future, radiologists will need to give much thought to whether it should be test package based or practice based (15). If the latter is the preferred route, our study indicates that it is feasible. Clearly the task is difficult, but it is not impossible (17). As noted by Brook et al (18), "First, it will never be possible to produce an error-free measure of the quality of care." However, this is no reason to avoid the exercise, and the results, if not the same, should clearly be congruent (19).

REFERENCES

1. Kopans DB. The accuracy of mammographic interpretation (editorial). N Engl J Med 1994; 331:1521-1522.
2. American College of Radiology. Wisconsin Medical Examining Board decision LS9310115MED Oct 26, 1995. In: ACR Bulletin (newsletter). Reston, Va: American College of Radiology, December 1995; 51:10-11.
3. American College of Radiology. Quality improvement manual. Section D-9. Reston, Va: American College of Radiology, 1990; 17.
4. American College of Radiology. Medical imaging privileges and peer review. In: A guide to continuous quality improvement in medical imaging. Reston, Va: American College of Radiology, 1996; 14-15.
5. Snedecor GW, Cochran WG. Statistical methods. Ames, Iowa: Iowa State University Press, 1967.
6. Garland LH. Studies on the accuracy of diagnostic procedures. AJR 1959; 82:25-33.
7. Herman PG, Hessel SJ. Accuracy and its relationship to experience in the interpretation of chest radiographs. Invest Radiol 1975; 10:62-67.
8. Lehr JL, Lodwick GS, Farrell C, Braaten MO, Virtama P, Kolvisto EL. Direct measurement of the effect of film miniaturization on diagnostic accuracy. Radiology 1976; 118:257-263.
9. Rhea JT, Potsaid MS, DeLuca SA. Errors of interpretation as elicited by a quality audit of an emergency radiology facility. Radiology 1979; 132:277-280.
10. Renfrew DL, 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; 183:145-150.
11. Berlin L, Berlin JW. Malpractice and radiologists in Cook County, IL: trends in 20 years of litigation. AJR 1995; 165:781-788.
12. Kundel HL. Perception errors in chest radiography. Semin Respir Med 1989; 10:203-210.
13. Hopper KD, Rosetti GF, Edmiston RB, et al. Diagnostic radiology peer review: a method inclusive of all interpreters of radiographic examinations regardless of specialty. Radiology 1991; 180:557-561.
14. Linver MN, Osuch JR, Brenner RJ, Smith RA. The mammography audit: a primer for the Mammography Quality Standards Act (MQSA). AJR 1995; 165:19-25.
15. Friedman DP, Rosetti GF, Flanders AE, et al. MR imaging: quality assessment method and ratings at 33 centers. Radiology 1995; 196:219-226.
16. Carey RM, Wheby MS, Reynolds RE. Evaluating faculty clinical excellence in the academic health science center. Acad Med 1993; 68:813-817.
17. Cascade PN. Quality improvement in diagnostic radiology. AJR 1990; 154:1117-1120.
18. Brook RH, McGlynn EA, Cleary PD. Quality of health care. II. Measuring quality of care. N Engl J Med 1996; 335:966-970.
19. Epstein A. Performance reports on quality: prototypes, problems, and prospects (sounding board). N Engl J Med 1995; 333:57-61.