Perspective

Simple Steps for Improving Multiple-Reader Studies in Radiology

Nancy A. Obuchowski[1] and Richard C. Zepp

Multiple-reader study designs have become popular in the radiology literature. We reviewed the major papers published in the *American Journal of Roentgenology* in the first 4 months of each of the years 1990 and 1995. The review was restricted to prospective studies of image interpretation. In the 1990 literature, we noted eight multiple-reader and 18 single-reader studies; in contrast, in the 1995 literature, we found 29 multiple-reader and eight single-reader studies. This trend reflects an increased awareness of the importance of multiple-reader studies. We examined the Results sections of the 29 multiple-reader studies from 1995 to assess the authors' motives for incorporating such a design. In 16 studies (55%), readers independently interpreted all images. However, the authors usually reported only the average interpretation of the readers; in only seven of the 29 studies (24%) did the authors describe differences among readers' interpretations. In 13 studies, interpretations were performed exclusively through "consensus reading." The method(s) used to achieve a consensus often were not explained. Only two of the 29 studies had more than three readers. In contrast, all of these studies included multiple patients. The average patient sample size was 45. Furthermore, differences observed among patients were routinely reported and/or depicted.

What Is Wrong With Our Multiple-Reader Studies?

Many published works on the diagnostic ability of an imaging technique suffer from an overly simplistic approach to image interpretation. Images are interpreted by a few (expert) readers, and any differences in interpretation are lost in the "consensus" result. In these studies, the focus is on the machine. Beam et al. [1] describe the readers in such studies as merely "repeated measurements of the diagnostic ability of [the] imaging machines."

However, imaging machines do not make diagnoses; radiologists, with the aid of imaging machines, make diagnoses. Beam et al. [1] wrote: "The focus in diagnostic radiology research must ultimately be on measuring and comparing the extent to which imaging technologies enhance the ability of radiologists to diagnose. From this perspective, imaging techniques are akin to 'treatments' that we apply to subjects (radiologists) and the responses we measure in these subjects are diagnostic success rates." To take this analogy further, consider how little we would learn about a "treatment" if the study had only one subject (single-reader study) or if the study had multiple subjects but their individual responses were not recorded but rather were expressed as a single pooled response (consensus reading study).

Given that researchers share this latter perspective on the significance of readers in image interpretation and assuming that this perspective is what has motivated the majority of authors of recent studies to gather responses from multiple readers, the next step to improving this research will be nearly painless.

Primary Objective of the Multiple-Reader Study

The motive for including multiple readers in a study is identical to the motive for including multiple patients in a study: inherent diversity. Patients manifest diversity in anatomic and imaging characteristics. Readers have disparate visual and cognitive abilities that lead to differences in interpretation.

Received August 1, 1995; accepted after revision October 2, 1995.

[1]Both authors: Department of Biostatistics and Epidemiology and Department of Radiology, The Cleveland Clinic Foundation, 9500 Euclid Ave., Cleveland, OH 44195-5196. Address correspondence to N. A. Obuchowski.

AJR 1996;166:517–521 0361–803X/96/1663–517

Given this diversity in readers' interpretations, the multiple-reader study design has four objectives: (1) to determine the frequency with which readers interpret the same case differently, (2) to assess the magnitude of differences in the interpretation of the same case, (3) to characterize differences in readers' overall performances, and (4) to estimate the performance of the "average" reader. The primary objective, however, is the last of these four objectives. We use the fictitious data in Table 1 to illustrate the idea of the average reader.

A technique is often described as having a particular accuracy. The accuracy of the technique, however, is not the same as the accuracy of any single reader. The accuracy of the technique is best described by the average accuracy of all readers in a particular population. We can estimate the average accuracy of a particular population of readers by studying a sample of readers from that population. In Table 1, four readers assigned diagnoses to five disease-positive patients with each of two competing techniques, *x* and *y*. Reader 1 is designated the senior reader. The performance of technique *x* is best estimated by the average performance of the four readers in this study, namely, a sensitivity equal to 65%; similarly, the estimated average performance of technique *y* is 65%. From this study of average performance, we would conclude that the accuracies of techniques *x* and *y* for the average reader are similar. However, we would want to test formally for differences in the performance of an average reader. Several statistical methods [2–5] are available for comparing average performances between techniques; we discuss these later.

The number of readers is one important consideration in designing a multiple-reader study. With a sample of two readers, one can compute an estimate of the average performance, but estimates based on a sample size of two are not reliable in describing a population. Methods for determining a required reader sample size for a study have been proposed [3] and are discussed later.

Another important consideration in designing a multiple-reader study is the representativeness of the reader sample. The applicability (or generalizability) of the results of a study is limited to the representativeness of the sample. Results from studies involving expert readers from a single institution are only applicable to expert readers from that institution. Reader populations broader than those currently being used (i.e., multiple subspecialty, multiple institution, and academia or private practice readers with a wider spectrum of experiences) will be required to fully characterize the performance of imaging systems. Beam et al. [1] have proposed a strategy for conducting such studies.

Limitations of Consensus Readings

To achieve the primary objective of multiple-reader studies, readers must interpret and report their findings independently of one another. Such independent reporting does not occur with consensus readings. Sometimes, a consensus reading represents the majority interpretation of the readers present; however, variability in the readers' interpretations is never recorded—only a single consensus score is. Other times, a consensus reading represents a biased (weighted) average, for example, when junior readers are influenced by senior readers; in this scenario, it is difficult to generalize findings on the basis of a single consensus score.

Reasons for the obsession of radiologists with consensus readings include an inappropriate focus on the accuracy of the imaging system rather than on the accuracy of the readers interpreting the images, the lack of robust statistical methods for dealing with complex multiple-reader data sets, and the ease of conducting and analyzing such studies. However, for the results of research studies to have any practical application, they must imitate the normal activities of radiologists and not a peculiar arrangement rarely seen in ordinary practice.

The data in Table 1 can be used to illustrate the ramifications of a consensus reading. Suppose that the consensus reading is truly the majority opinion of the readers present. The data from this consensus reading might look like the data in the "majority" columns in Table 2. We could define the performance of the techniques in terms of the performance of the consensus reading with each technique. How-

TABLE 1: Fictitious Test Results for Tumor Detection in Five Patients with Confirmed Disease

Patient and Sensitivity	Score Assigned by the Indicated Reader with the Following Technique:							
	x				*y*			
	1	2	3	4	1	2	3	4
Patient								
1	2	2	2	2	2	2	1	2
2	1	2	1	1	1	1	0	1
3	0	1	1	1	2	0	1	0
4	2	0	0	0	2	0	0	0
5	0	0	0	1	1	1	0	1
Sensitivity[a]	60	60	60	80	100	60	40	60

Note.—Scores assigned by individual readers were as follows: 0 = absent, 1 = suspicious, 2 = present. Reader 1 is the senior reader.

[a]Sensitivity is defined as the percentage of patients assigned a score of 1 or 2. Average sensitivities were 65% for both techniques.

TABLE 2: Test Results from Two Types of Consensus Readings Based on Data in Table 1

Patient and Sensitivity	Score Resulting from the Indicated Type of Consensus Reading with the Following Technique:			
	x		*y*	
	Majority	Weighted	Majority	Weighted
Patient				
1	2	2	2	2
2	1	1	1	1
3	1	1	0	2
4	0	2	0	2
5	0	0	1	1
Sensitivity[a]	60	80	60	100

Note.—Scores were as follows: 0 = lesion absent, 1 = suspicious lesion, 2 = lesion present.

[a]Sensitivity is defined as the percentage of patients assigned a score of 1 or 2.

ever, the performance of the consensus reading (sensitivity of 60% for both techniques) will not necessarily match the average performance of the readers in the sample (65% for both techniques; see Table 1).

Now consider the effect of a consensus reading in which one reader influences the test results of the other readers. Plausible data for this weighted consensus reading, based on the data in Table 1, are given in the "weighted" columns in Table 2. The consensus scores are similar to the test results of reader 1, the senior reader. This weighted consensus reading yields biased performance estimates. Specifically, the data from such a study would incorrectly suggest that reader performance with technique *y* is superior to that with technique *x*. It is clear that this suggestion is a reflection of one of the four readers and not an accurate reflection of the effect of technique *y* on the average reader.

Other Objectives of the Multiple-Reader Study

Other objectives of the multiple-reader study focus on characterizing variability in readers' interpretations. These objectives are described below, along with simple methods to achieve them.

Determination of the Frequency with Which Readers Interpret the Same Case Differently (Objective 1)

When readers in a study independently evaluate the same sample of images, readers' interpretations for each case can be compared. A useful measure of the frequency with which readers interpret the same case differently is the percentage of cases in which there is disagreement. Consider the data in Table 1 for technique *x*. There is some level of disagreement in interpretation for 80% (four of five) of the patients (i.e., patients 2–5); there are contradictory interpretations for 20% (one of five) of the patients (i.e., patient 4).

More sophisticated analyses, such as unweighted kappa statistics [6], also are applicable. Kappa statistics describe the amount of agreement between two readers, over and above chance agreement. Consider, for example, readers 3 and 4 using technique *x* in Table 1. Their diagnoses agree 80% of the time. However, by chance these two readers would be expected to make the same diagnosis occasionally (specifically, 36% of the time). The kappa statistic is the observed agreement (80%) minus the chance agreement (36%), the result of which is divided by the amount of non-chance agreement possible (100% – 36% = 64%). For this example, kappa equals 0.69. Kappa values often are interpreted as follows: ≤0.00 = poor agreement, 0.01–0.20 = slight agreement, 0.21–0.40 = fair agreement, 0.41–0.60 = moderate agreement, 0.61–0.80 = substantial agreement, and 0.81–1.00 = almost perfect agreement [7]. Thus, readers 3 and 4 demonstrate substantial, but less than perfect, agreement.

Assessment of the Magnitude of Differences in the Interpretation of the Same Case (Objective 2)

Once differences in readers' interpretations are noted and their frequency is reported, the magnitude of these differences becomes important. For each case, the range of interpretations (i.e., the difference between the two extreme interpretations) can be computed. The average range, along with notable cases of reader discordance, could be reported. In the example shown in Table 1 for technique *x*, the ranges of scores for the five patients are 0, 1, 1, 2, and 1; the average range is 1. Patient 4 is notable: three readers reported the absence of a tumor, whereas the other reader, reader 1, designated the senior reader, reported the presence of a tumor.

A more sophisticated approach might involve a weighted kappa statistic [6]. This version of the kappa statistic gives partial credit for some types of disagreement. For example, the agreement between test scores of 0 (absent) and 2 (present) would be scored as 0, the agreement between test scores of 0 (absent) and 0 (absent) would be scored as 1, and the agreement between test scores of 0 (absent) and 1 (suspicious) would be scored as 0.5, even though, in the purest sense, there is no agreement in this case. For readers 3 and 4 using technique *x* (Table 1), the weighted kappa statistic is 0.74. The weighted kappa statistic is slightly higher than the unweighted kappa statistic because of the partial credit given for patient 5.

Characterization of Differences in Readers' Overall Performances (Objective 3)

It is possible for readers to interpret individual cases differently but to have similar performances over a sample of cases. To characterize readers' overall performances, an appropriate summary measure of performance must be identified. Common measures of overall performance in radiology studies include sensitivity and specificity (for accuracy studies in which the test results are classified as "positive" or "negative"), the area under the receiver operating characteristic curve [8] (for accuracy studies in which the test results have multiple possibilities and in which the possibilities can be ordered from low to high probability of disease), and a simple average or median (as in a summary of tumor size or image quality scores).

Sensitivities were computed for each reader using both techniques in Table 1. Note that for readers 1, 2, and 3 using technique *x*, the overall sensitivities were the same, even though these readers disagreed on diagnoses for individual cases. Compare the overall test results obtained with the two techniques for each reader. Reader 1 performed best using technique *y*, whereas readers 3 and 4 performed best using technique *x*. However, such observations should be followed by an appropriate statistical test to assess the significance of any observed differences (i.e., McNemar's test [9] when the summary measure is sensitivity or specificity, a *z*-test [10] for the area under the receiver operating characteristic curve, a paired *t*-test for averages, or Wilcoxon's signed rank test for medians). Reader-specific results are important because they describe the differences between readers' diagnostic abilities and the different effects of imaging systems on different readers.

Objectives 1 through 3 cannot be achieved when consensus readings are used. From the data in Table 2, we cannot evaluate the frequency with which readers' interpretations

differed for the same case (objective 1), and we cannot assess the magnitude of any differences (objective 2); we also cannot examine differences in readers' performances (objective 3).

State-of-the-Art Statistical Methods for Multiple-Reader Studies

Although the simple methods described above are adequate for describing reader differences in image interpretation, more sophisticated methods are required for testing hypotheses about the performance of an average reader. Several statistical methods for analyzing multiple-reader studies of diagnostic accuracy recently were described [2–5]. These methods can be used to answer the following questions about the performance of imaging systems. (1) Which of several competing techniques has greater average performance, and if competing techniques have similar average performances, which technique is associated with less diagnostic variability? (2) What is the range of performance of an imaging system for all patients and readers? (Note that this range of performance is not the same as the range of performance for a particular sample of patients and readers. This question relates to the range of performance for the entire population of patients and readers.) (3) How much of the total variability in performance is attributable to differences among readers and differences among patients? (4) Are there subpopulations of patients and/or readers for which the imaging system demonstrates exceptionally high or low performance?

Table 3 provides a brief summary of four statistical approaches to multiple-reader studies of diagnostic accuracy. Jackknife describes the method of Dorfman et al. [2], in which a resampling technique (called jackknifing) is used to estimate the variability in the data. Corrected *F*-test describes the method of Obuchowski [3], in which the test statistic (*F*-statistic) from a usual analysis of variance is corrected for the correlations between observations. The ordinal regression approaches of Toledano and Gatsonis [4] and Gatsonis [5] use ordered (ordinal) test results, such as 1 = definitely normal, 2 = probably normal, 3 = possibly abnormal, 4 = probably abnormal, and 5 = definitely abnormal. A model is then fit to explain the relationship between the test scores and the factors that affect those scores (e.g., disease status, reader, and technique). The ordinal regression approaches thus fit a model to the actual test results. The corrected *F*-test and jackknife methods, on the other hand, fit a model to a summary measure (for example, the area under the receiver operating characteristic curve, which is a summary measure of the test results) or to its resampled counterpart (pseudovalue of summary measure). Because the ordinal regression approaches fit a model to the test results instead of to a summary measure, these approaches are more flexible and readily incorporate additional information about patients and readers (e.g., patient covariates, such as age, gender, and weight, and reader covariates, such as number of years of experience, subspecialty, and special training). These approaches, however, require validity checks.

Several concepts must be appreciated when these statistical methods are applied. One of the key issues is accountability for the lack of independence in readers' test results. Even though readers interpret images independently of one another, they all interpret the same images, a fact that introduces correlation. All four methods shown in Table 3 take this correlation into account.

Another important statistical issue is the manner in which the model characterizes readers. Ideally, the model should depict readers in a sample as representing readers from the population; statisticians refer to such a model as the "random reader effect" model. The results of a random reader effect model can be generalized to the population of readers represented by the sample. Alternatively, a model can be chosen so that specific readers in a study are its focus; this type of model is known as the "fixed reader effect" model. Because a fixed reader effect model only focuses on readers in the study, conclusions from an analysis cannot be generalized to an entire population of readers. Thus, the random reader effect model has broader applicability than the fixed reader effect model.

Last, a strategy for determining required sample sizes (for both patients and readers) on the basis of the corrected *F*-test approach has been developed [3]. This method also has

TABLE 3: State-of-the-Art Statistical Methods for Multiple-Reader Studies

Method	Model	Advantages	Disadvantages
Jackknife [2]	Pseudovalue of summary measure	Uses well-known ANOVA setting	Assumes linear, additive model
Corrected *F*-test [3]	Summary measure	Uses existing software; power analysis available	Assumes linear, additive model; more rigid correlation structure
Ordinal regression			
Fixed effects [4]	Test results	Flexible; can incorporate covariates	Readers treated as fixed effects
Random effects [5]	Test results	Flexible; can incorporate covariates	Computationally intense

Note.—ANOVA = analysis of variance.

been used to investigate the power of a variety of multiple-reader study designs [11].

Conclusion

Although multiple-reader studies are prevalent in radiology research, their potential for providing valuable information about a technique has been neglected. An important goal of multiple-reader studies is to assess the variability in interpretations attributable to reader differences. Consensus readings miss this goal completely. Simple changes are needed in the way multiple-reader studies are conducted, presented, and analyzed. These changes can be summarized: avoid consensus readings, report notable differences in readers' interpretations of the same case, and present reader-specific summary measures. Sophisticated statistical methods are available to answer complex questions about the performance of readers using competing techniques.

ACKNOWLEDGMENTS

We thank Mark E. Baker for suggesting the topic of this manuscript and for many insightful discussions. We also thank the reviewers, whose comments improved the presentation of the manuscript.

REFERENCES

1. Beam CA, Baker ME, Paine SS, Sostman HD, Sullivan DC. Answering unanswered questions: proposal for a shared resource in clinical diagnostic radiology research. *Radiology* **1992**;183:619–620
2. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. *Invest Radiol* **1992**;27:723–731
3. Obuchowski NA. Multireader, multimodality receiver operating characteristic curve studies: hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations. *Acad Radiol* **1995**;2[suppl 1]:S22–S29
4. Toledano A, Gatsonis CA. Regression analysis of correlated receiver operating characteristic data. *Acad Radiol* **1995**;2[suppl 1]:S30–S36
5. Gatsonis CA. Random-effects models for diagnostic accuracy data. *Acad Radiol* **1995**;2[suppl 1]:S14–S21
6. Fleiss JL. *Statistical methods for rates and proportions*, 2nd ed. New York: Wiley, **1981**
7. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics* **1977**;33:159–174
8. Metz CE. ROC methodology in radiologic imaging. *Invest Radiol* **1986**; 21:720–733
9. Dwyer AJ. Matchmaking and McNemar in the comparison of diagnostic modalities. *Radiology* **1991**;178:328–330
10. Hanley JA, McNeil BJ. Comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology* **1983**; 148:839–843
11. Obuchowski NA. Multireader ROC studies: a comparison of study designs. *Acad Radiol* **1995**;2:709–716