Measuring Observer Performance in Chest Radiology: Some Experiences

E. James Potchen, MD

All decisions made under conditions of uncertainty have error rates. All meaningful decisions are made under conditions of uncertainty. Can this uncertainty be measured? Can variations in how different observers deal with this uncertainty be ascertained? The ability to measure observer performance in diagnostic imaging was one of the issues that initiated the field of medical decision analysis. This article exemplifies an approach and is worth discussing as a preamble to presenting our long-term project of measuring variations in observer performance. The paper focuses on the interpretation of chest x-ray images, although the principles and findings described can be applied to nearly every radiologic modality and interpretation task.

Key Words: Medical decision making, observer performance, chest imaging

J Am Coll Radiol 2006;3:423-432. Copyright © 2006 American College of Radiology

INTRODUCTION

The pioneering work of Lee Lusted and Eugene Saenger, who were among those who founded the Society for Medical Decision Analysis, foretold a remarkable opportunity to better understand variation in human decision making on the basis of how different individuals observe and interpret radiographic images.

Over the years, many techniques have been developed to evaluate how different observers reach conclusions when interpreting a radiographic image. Observer performance studies have been used in a wide variety of medical imaging research, with more than 200 articles published in recent years. Many recent papers have shown the range of applications for these observer performance studies [1-14]. One recent article is particularly useful in describing the utility of observer performance measurements. Shah et al [15] evaluated the merits of alternative ways to review images obtained with modern imaging modalities. They studied the effect of a computer-aided diagnosis (CAD) system when used to detect and diagnose solitary pulmonary nodules. The present article exemplifies an approach and is worth discussing as a preamble to presenting our long-term project of measuring variations in observer performance.

Shah et al [15] appraised the effect of different levels of experience in distinguishing between benign and malignant solitary pulmonary nodules on computed tomography (CT). They studied 3 different interpretation conditions: (1) when only image data were presented, (2) with the addition of clinical data, and (3) with the use of a CAD system. Shah et al [15] used 28 thin-section CT data sets with proven diagnoses (15 malignant and 13 benign) and asked each observer to assign a level of confidence from 0.0 to 1.0, where 0.0 was benign and 1.0 was malignant. They repeated these observations for each of the 3 conditions. The performance metric they used was a multiple-reader, multiple-case receiver operating characteristic (ROC) analysis. Shah et al [15] used a variety of observers: 1 thoracic radiology fellow, 2 non-thoracic radiologists, 3 radiology residents, and 3 thoracic radiologists. The average areas under the ROC curves for all observers at each stage were 0.68, 0.75, and 0.81 for image data alone, with clinical data, and with the CAD system, respectively. The differences in performance were statistically significant. On the basis of these data, Shah et al [15] concluded that the addition of CAD made a significant improvement in the diagnosis of solitary pulmonary nodules.

For many years, my group has been studying observer performance in chest radiology [16]. We have shown a standard set of posterior-anterior chest x-rays to more than 100 radiologists from different radiology groups in different areas of the world. We observe how different individuals make observations and interpret films. We have found that if individuals are informed of how they vary from the norm, they can, and at times do, improve the quality of their diagnostic interpretations. Thus, the measurement of observer performance can be a tool used to improve the diagnostic accuracy of radiologists in reading chest x-rays. Because we have not studied images more complex than chest x-rays, we do not know how

Department of Radiology, Michigan State University, East Lansing, Mich.

Corresponding author and reprints: E. James Potchen, MD, Michigan State University, Department of Radiology, 160 Radiology Building, East Lansing, MI 48824-1313; e-mail: jim.potchen@radiology.msu.edu.

0091-2182/06/$32.00 • DOI 10.1016/j.jacr.2006.02.020

Table 1. Six steps in the value chain of diagnostic imaging

1. Selection of the patient and the appropriate procedure
2. Generating the image
3. Observing the image
4. Interpreting the observation
5. Communicating the interpretation
6. Using the information to benefit the patient

this type of assessment would apply in more complicated image data sets, such as the multiple images found in modern-day CT or magnetic resonance. However, an appreciation of how to study observer performance in a relatively simple data set, such as a series of chest x-rays, may aid in understanding more sophisticated approaches to assessing observer performance with much larger data sets, as are found in the traditional radiologic practices of today. My group has compared and contrasted radiologists' performance in different geographic centers, in different academic or private practice settings, and with different levels of experience in interpreting radiologic films. We have made a concerted effort to understand the marginal utility of having learned radiology or what in the process of learning radiology makes a difference in the interpretive skills of an observer. We have primarily sought to develop and test tools that will allow radiologists to compare their performance against standards set by other radiologists' performance when faced with making the same decisions. This paper reviews some of this experience.

DIAGNOSTIC IMAGING VALUE CHAIN

Diagnostic radiology is an important component of the clinical information system in patient care. Information is defined as a reduction in uncertainty. The purpose of any diagnostic procedure is to diminish clinical uncertainty. Although I have emphasized observer performance as a component of the chain of value added by radiology, I do not mean to lessen the importance of other aspects of diagnostic radiology in adding value through the radiologic process. The chain of value in diagnostic imaging, as outlined in Table 1, begins with the selection of a patient and an appropriate procedure to address the uncertainty that is present in a specific clinical situation. An image is then generated, and this image is observed and interpreted. The observer then reaches some conclusion that is communicated to the referring physician, who must use this information to benefit the patient before value can be added.

An observer performance measure could be based on the ability to detect an abnormality or the decisions made once an abnormality is detected. In understanding the chain of value added in the process of diagnostic imaging, one cannot rely merely on the detection and recognition of an abnormality. For a diagnostic procedure to add value, the information it obtains must be communicated to someone who will use it to help the patient. The entire sequence warrants monitoring, and the observer performance study is but one component in this chain of value in diagnostic imaging. In my group's studies, we have found that the variation in communication is at times as great as, if not greater than, the variation in the performance of the observer [17].

Information is defined as a decreased randomness in the state of knowledge. It can be measured using Shannon's [18] neg-entropy, which essentially measures the amount of randomness in any given information set. A diminished randomness (whereby more order is put into some disordered system) results in increased information. Thus, information is decreased randomness in the state of knowledge, and neg-entropy is a measure of that information. Quality improvement in diagnostic imaging depends in part on decreasing variance in the performance of the involved professionals. If we can measure how well observers can perform, we can set benchmarks against which multiple observers can be compared.

INTRAOBSERVER DISAGREEMENT

Intraobserver disagreement has been an issue in obtaining reproducible results from observer performance measurement [3]. How important is this problem? What can be done to improve observer consistency? My group studied a randomized set of 60 chest x-rays, asking radiologists to sort them on the basis of what they observed on the films. Initially, we asked them to separate the films into a group of "normal" films and a group of "abnormal" films. Individual observers were not consistent in their use of these words. We found wide variation in what the words *normal* and *abnormal* were interpreted to mean. Is "abnormal" something a radiologist does not usually see? Or is it something that is 2 standard deviations from the "norm"? Is it something that is clinically significant? We then repeated the study, asking the radiologists to separate the films in response to the question, "Is there anything on this film which, if not detected and reported, would adversely affect this patient?" This is a standard question that is asked to determine whether malpractice has occurred. Error alone is not malpractice, and the simple fact that errors are made is not tantamount to legal liability. To reach the threshold required for successful malpractice litigation, there must be something clinically significant on a film that, if not reported, would harm the patient.

This clinical impression as a metric has more relevance

Table 2. Diagnostic impressions

1. Normal
2. Abnormal—but not clinically significant
3. I'm not certain—warrants further diagnostic study
4. Abnormal—diagnosis uncertain—warrants further diagnostic study
5. Abnormal—diagnosis apparent—warrants appropriate clinical management

in my group's studies than merely the radiologic call of normal or abnormal. This distinction revealed some interesting intraobserver observations. For example, in the series of 60 chest x-rays (with no patient identifiers), we repeated some of the films. When radiologists were asked, "Is the film normal?" they disagreed among themselves in how they classified the same films an average of 20% (range 5% to 30%) of the time. When they were asked, "Is there anything on the film which, if not identified, would adversely affect the patient?" the range of intraobserver disagreement decreased to a mean of 7% (range 3% to 21%).

Thus, in measuring observer performance, it is important to minimize the variance in the mind of observers when performing a study that is induced by a lack of clarity about what is expected. A radiologist should not merely be asked whether a film is normal or abnormal; a different metric is needed. In view of this, my group then moved to using a 5-point scale to elucidate diagnostic impressions, similar to what is used in traditional ROC analysis [19-22]. We have sought a reproducible technique to minimize intraobserver variation on repeated studies by the same individual. We have found that a scale of graded choices (Table 2) diminishes intraobserver inconsistencies. From this, we conclude that changing the observer performance study protocol to focus on patient care increases an interpreter's consistency in reporting what he or she observes on films.

THE SAGA OF THE MISSING LEFT MEDIAL CLAVICLE

Many years ago, my group began using a film of a missing left medial clavicle that was obtained from the ACR's teaching file data set (Figure 1). Many radiologists had previously seen it and were aware of the difficulty in making this interpretation. However, we found this to be an excellent film to measure the difference in an individual's ability to realize what was not on the film rather than merely looking at a positive density in the chest. We included this in the set of 60 chest films. In an initial study, when no cues were given, 60% of the observers failed to identify this abnormality. We then repeated the study with different radiologists as observers. When we used the erroneous cue of the clinical indication being an "annual physical examination," 58% of our observers missed the abnormality and placed this film in the group of normal films. However, when the correct clinical indicator, pertaining to metastatic survey, was added, 83% of the observers found the abnormality, indicating that in this case, the correct clinical cue substantially increased the fidelity of the observer's performance.

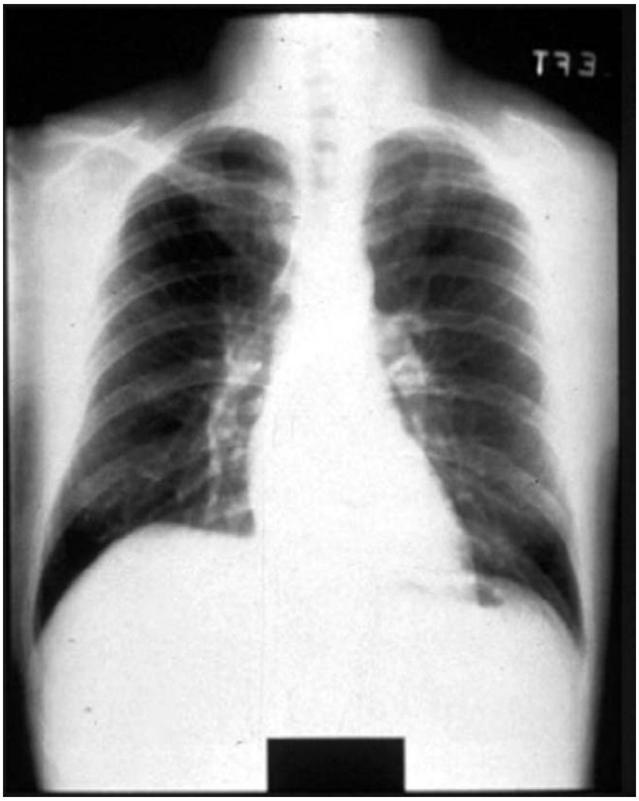

Fig 1. Missing left medial clavicle.

THE EFFECTS OF CLINICAL CUES ON INTERPRETATION

"People see what they are prepared to see," said Ralph Waldo Emerson. My group has observed that the effect of clinical cues depends in part on the difficulty of diagnosis. To separate the films by degree of difficulty, we ranked the data set into 4 groups on the basis of the probability that observers would reach correct conclusions. We termed these 4 classes of films obvious normal, difficult normal, difficult abnormal, and obvious abnormal. By classifying the individual films in this way, we were better able to appreciate the effect of direct, irrelevant cues on observers' performance (Table 3). Correct and irrelevant cues, or no cue, made very little difference to an observer's conclusion that a film was normal in the group of films classed as normal. In the difficult-normal

Table 3. Proportion of correct calls by difficulty of diagnosis and relevance of clinical cue (complaint)

Degree of Difficulty Based on the Diagnosis	Directive (Correct) Cue	Irrelevant Cue	No Cue
Normal (obvious)	0.81	0.78	0.81
Normal (difficult)	0.73	0.73	0.73
Abnormal (difficult)	0.67	0.48	0.44
Abnormal (obvious)	0.89	0.82	0.92

class, it made virtually no difference whether we presented a correct, irrelevant, or no cue to an observer. However, for those films classed as difficult abnormal (ie, abnormal films that were often called normal by observers), we found that a direct cue had little effect on the interpretation, but an irrelevant or no cue diminished the ability to detect the abnormality. On obvious abnormal films, there was no significant difference in the effect of the cue. Thus, in our experience, whether clinical cues changed diagnostic accuracy in observer performance studies depended in part on the difficulty of the film as measured by variations of how well observers agreed in the conclusion reached. If we provided a correct clinical history, we increased the detection of difficult abnormalities but had relatively little effect on other types of images under study.

COMPARING INDIVIDUAL FILMS: THE EFFECT OF FILM SELECTION ON OBSERVER PERFORMANCE

Gur et al [23] studied the prevalence effect in a laboratory environment. They found no significant effect as a function of prevalence for any abnormality, group of cases, or readers. They concluded that under such laboratory conditions, a prevalence effect exists, but "it is quite small in magnitude and will not likely alter the conclusions derived from such studies."

My group's observers were asked to rate the films in the previously described categories. From this, we were able to calculate a number of indices that characterized the variations in individual observer performance (Table 4). Each film can be studied to measure its impact on the set of films under study. For example, Table 5 displays some of the results when 60 films were submitted to a group of 30 radiologists. "Percentage correct" is the frequency with which a normal film was rated 1 or an abnormal film was rated 4 or 5. A false-negative film was defined as an abnormal film that was rated 1. A false-positive film was defined as a film that was normal but rated 4 or 5. "Percentage equivocal" refers to films that were rated 3 or 4, for which a radiologist was not sufficiently confident to reach a conclusion of normality or abnormality. "Percentage additional studies" refers to any observations that required additional imaging examinations. These were the films that were rated 3 or 4 (ie, "I'm not certain—warrants further diagnostic study" or

Table 4. Indices of observer performance

Index	All Physicians (n = 128)	Board-Certified Radiologists (n = 95)	Radiology Residents (n = 12)	Non-radiologists (n = 21)
Receiver operating characteristic	82%	86%	80%	66%
Probability of true-positive[a]	78%	81%	83%	67%
Probability of true-negative[b]	65%	70%	64%	42%
Probability of false-negative[c]	14%	12%	16%	20%
Probability of false-positive[d]	31%	20%	31%	39%
Ambiguity level[e]	10%	9%	3%	16%
Confidence parameter[f]	1.18	1.17	1.08	1.15
Discrimination parameter[g]	1.43	1.67	1.22	0.61

[a] Total number of positive films rated 4 or 5/total number of positive films.

[b] Total number of negative films rated 1 or 2/total number of negative films.

[c] Total number of positive films rated 1 or 2/total number of positive films.

[d] Total number of negative films rated 4 or 5/total number of negative films.

[e] Total number of films rated 3/total number of films.

[f] Variance of negative film distribution/variance of positive film distribution (the confidence parameter should be 1.0 to minimize error).

[g] Mean of positive film distribution–mean of negative film distribution/variance of positive film distribution (the discrimination parameter should be maximized for optimum discrimination capacity).

Table 5. Observer performance per film (60 films reviewed by 30 radiologists; definitions in text)

Film Number	% Correct	% False-Negative	% False-Positive	% Equivocal	% Additional Studies
1	64.71	29.41	—	5.88	5.88
4	52.94	—	35.29	11.76	47.05
5	64.70	29.41	—	5.88	5.88
8	29.41	—	58.83	11.76	52.94
12	35.29	52.94	—	11.76	11.76
18	64.71	23.53	—	11.76	11.76
27	52.94	—	41.18	5.88	29.41
29	52.94	41.18	—	5.88	5.88
33	52.94	—	35.29	11.76	23.52
35	41.18	52.94	—	5.88	5.88
59	29.41	—	47.08	23.53	47.06
60	52.94	—	41.17	5.88	11.76

"Abnormal—diagnosis uncertain—warrants further diagnostic study"). The number of additional films required is an important issue in that ambiguity in a report can create a substantial difference in the health care costs incurred when a diagnostic examination is requested. Some physicians are highly certain, whereas others have much more difficulty making a commitment with a level of certitude sufficient to be clinically useful.

COMPARING OBSERVERS

The underlying assumptions in studying these sets of films were that there were positive and negative films in the group being studied. The quantitative, categorical ranking described above provides for a binormal distribution that would have independent means and independent variances. Using these assumptions, one can envision a simple distribution of individual observers (Figure 2). My group has found this theoretical foundation helpful in communicating the nature of an observer performance study to individual observers. Individuals can then become aware of their decision threshold and thereby adjust their management of uncertainty to obtain the outcome they desire. This is a highly individualized and personalized approach to improving observer performance.

Receiver operating characteristic analysis is the primary method used to study variations in individual or group performance [21,24]. The value added by an observer can be seen in the area under an ROC curve. This is a measure of information added by the observer. There was remarkable variation in how much value was added by different radiologists when observing this standard set of films. The extremes of performance among board-certified radiologists were quite remarkable (Figure 3). In this set of observers, the most accurate observer was observer 16, with an accuracy of 0.97 and a *z* score 2.65 standard deviations above that of the average observer. This individual was remarkably skilled in interpreting this set of films. By the same token, another radiologist (observer 10) had an accuracy of 0.67 and a *z* score of 5 standard deviations less than that of the average observer. Using the area under the ROC curve as a measure of value added by an observer to a set of films, this range of performance indicates that different observers clearly added substantially different value to the set of x-rays. Using just the simple ROC curve, it is possible to measure the value added by an observer in reading a set of films.

INDICES OF OBSERVER PERFORMANCE

When my group initially sought to study the variations in observer performance, we set the proportion of normal

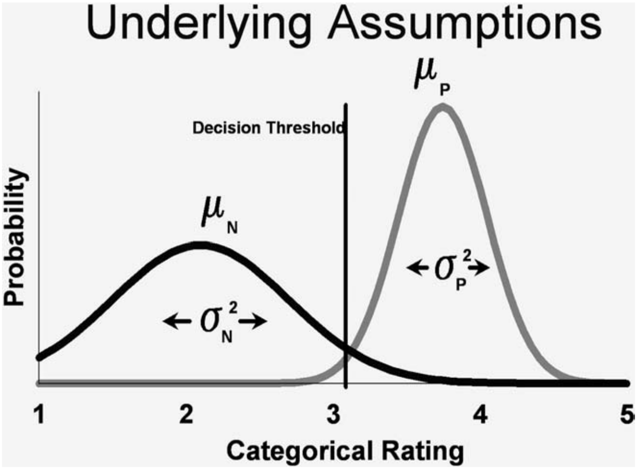

Fig 2. Binormal distribution (μ_N and μ_P, where μ is the independent mean, σ is the independent variance, N = negative, and P = positive).

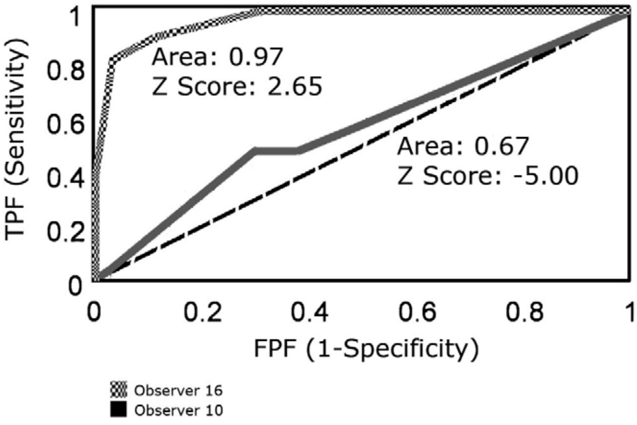

Fig 3. Receiver operating characteristic curves of the best and worst performers. FPF = false-positive films; TPF = true-positive films.

films to the average frequency of normal films in our practice. However, that approach provided so few abnormal films that we were unable to adequately assess the variation in individual performance for other than false-positive studies [25]. We therefore distributed the normal and abnormal films at approximately 50% each, although the observers were not informed of the exact number before observing the set of films. The observers were asked to rate the films in the previously described categories. From this, we were able to calculate a number of indices that characterized the variations in individual observer performance (Table 4).

Many measurable indices can be used to detect and analyze variations in observer performance in interpreting chest x-rays. These indices include false-positive rates, false-negative rates, true-positive rates, true-negative rates, and diagnostic accuracy as determined by the area under the ROC curve. In addition to these traditional metrics, my group has also made use of a discrimination parameter and a confidence parameter. Table 4 displays the results of one such study of 95 board-certified radiologists, 12 radiology residents, and 21 physicians who were not radiologists.

In comparing the performance differential between the best-performing radiologists and the worst-performing radiologists (Table 6) and between radiologists and nonradiologists, the predominant distinguishing feature was found in the discrimination parameter. The discrimination parameter measured the ability to discern normal from abnormal films. The difference between these groups in their discrimination parameter was considerably greater than the confidence parameter. This suggests that although poor performers may not be very accurate, they can be very confident. Observers' lack of ability to discriminate normal from abnormal films does not necessarily diminish their confidence (Table 6). My group has also measured the time it takes to read the set of films as an indication of observer decisiveness. Taken together, these variations in observer performance allow us to categorize radiologists and inform them of how they perform compared with other observers.

All observers have a characteristic way in which they manage the threshold of uncertainty in making decisions. Some people are risk takers, and they are likely to have more false-positive errors. Others are risk adverse, and they are more likely to have high false-negative rates. Still others cannot make up their minds, and they will have high ambiguity numbers and more frequently require additional films before reaching conclusions. Once observers are informed of how they compare with others who have seen the same set of films, they can adjust their thresholds in borderline situations by modifying their decision criteria. This allows for considerable individual correction. Most people are quite unaware of how they compare with others. One of the major reasons why radiologists have volunteered for this study is to be able to test how they perform compared with other radiologists in equivalent practices.

INDIVIDUAL ROCS

From these data, my group was able to supply individual radiologists with their ROC curves. This gave them an opportunity to modify their behavior. Some observers, when retested, demonstrated changes in their ROC curves by adjusting their threshold decisions to be more in conformity with the norm. Radiologists who initially had high false-negative rates did not significantly change their behavior when informed of this before retesting. It seems to be more difficult to correct observer performance for a high false-negative rate than for a high false-positive rate.

Table 6. Performance differential between the top 20 radiologists and the bottom 20 radiologists demonstrating the difference between the ability to discriminate and levels of confidence

	Diagnostic Accuracy (%)	Discrimination Parameter	Confidence Parameter
Top 20 radiologists	94.6	2.37	1.08
Bottom 20 radiologists	75	1.06	1.21

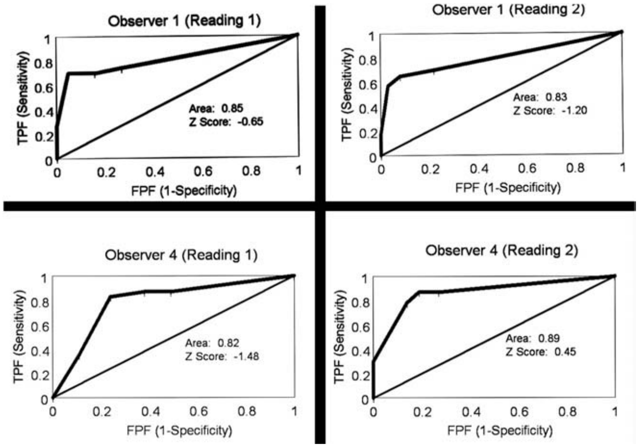

Fig 4. Consistencies in the shapes of the receiver operating characteristic curves in individuals when retested. FPF = false-positive films; TPF = true-positive films.

In retesting with different sets of films, my group noted considerable consistency in the shapes of the individual ROC curves. The shapes of the ROC curves and the overall accuracy were quite consistent when the test was repeated (Figure 4). For example, for observer 4 on the first reading, the area under the ROC curve was 0.82, with a *z* score of 1.48. On the second reading, the area under the ROC curve was 0.89. For observer 1, the area under the ROC curve was 0.85 on the first reading and 0.83 on the second reading. The shapes of these observers' ROC curves were quite different but consistent within observers. Most observers maintained the same patterns of their ROC curves on serial observations of their performance.

In addition to providing the observers with their unique ROC curves, my group found that individual performance measures rated by *z* scores allowed observers to better appreciate their decision-making profiles compared with those of other observers reviewing the same film set. This allowed individual observers an opportunity to adjust their decision thresholds to improve on their performance as observers of chest x-rays. Figure 5 represents a sample of these profiles furnished to observers.

Contrasting observers 1 and 3 shows that there was a significant difference between false-positives, with a minimal trade-off in false-negatives. In comparing observers 5 and 6 (Figure 6), one can see that observer 6 took longer to review the films and got more correct than the average observer. When these data are furnished anonymously to a group of radiologists, they can compare their performance with that of others in their group.

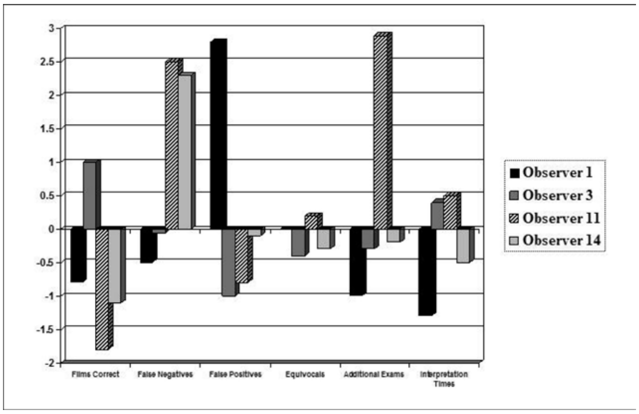

Fig 5. Comparison of 4 observers from one group of radiologists demonstrating remarkable difference in their decision-making profiles.

INTEROBSERVER VARIATIONS

These data also allow the presentation of an entire group's performance. Figure 7 represents one such group of radiologists. By having all members of the group know their individual performance as well as that of others in the group, the radiologists can better understand the range of behavior in their group as well as how their group compared with an external norm. These graphs were always furnished without attribution. Although anonymity maintained confidentiality, in many circumstances, radiology groups are quite capable of picking out which number represents which radiologist. Groups of radiologists often are aware of the decision-making profiles of the members of their group. They learn how their partners make decisions by working with them over many years. Often, however, an individual radiologist

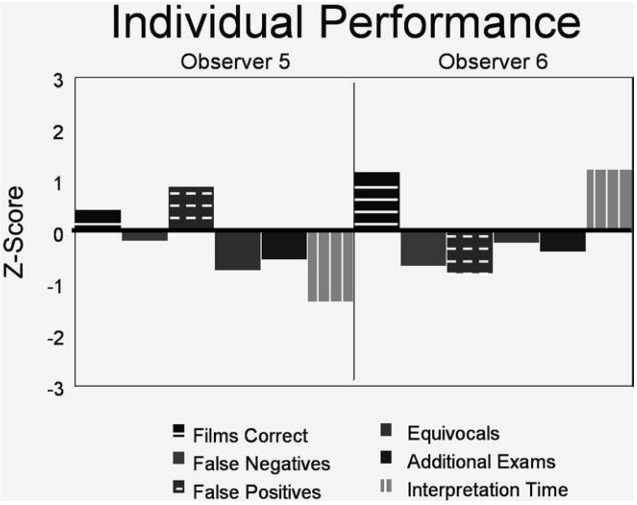

Fig 6. The *z* score represents the standard deviation from the norm.

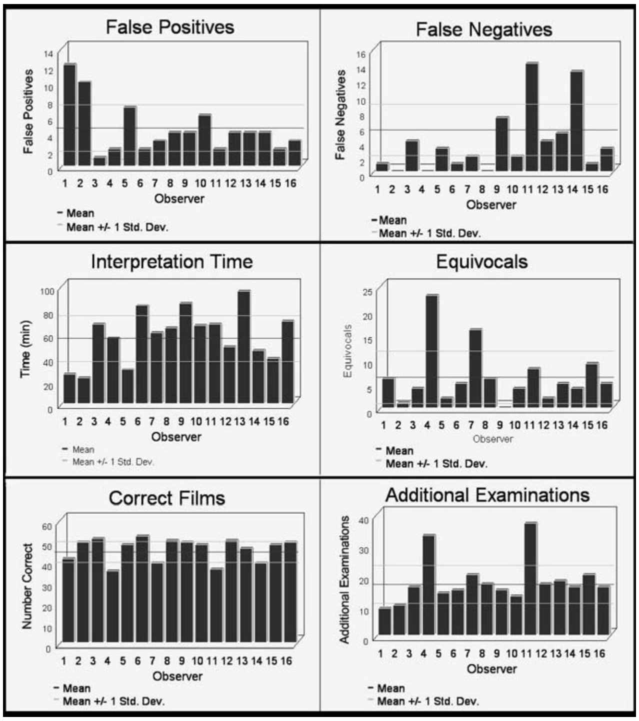

Fig 7. Variations within 16 members of one radiology group. Observers 4 and 11 required considerably more additional examinations than their partners in the same practice. Observers 1 and 2 had considerably more false-positives than their colleagues in the same group. Std. Dev. = standard deviation.

would have no prior appreciation of his or her own decision-making characteristics, even when he or she was making decisions in a manner that differed by more than 2 standard deviations from other radiologists within the group. This information obtained through self-examination allowed individual radiologists to modify their performance. Not all individuals corrected their performance when they were made aware of it, but many did, therefore shifting the norm of the group under study.

GEOGRAPHIC PERFORMANCE VARIATIONS

My group studied the variation in performance by geographic site of practice to determine if we could observe a difference in the behavior of different groups of radiologists. Figure 8 displays a comparison of performance in 7 different geographic locations. Although groups of radiologists in different geographic sites tend to demonstrate similar patterns in their decision making, some differences are demonstrable. Most geographic locations are not statistically distinguishable from other locations. There was, however, an exception seen in the local tolerance for equivocation. Site E had the lowest equivocation rate, whereas site A had the highest equivocation rate.

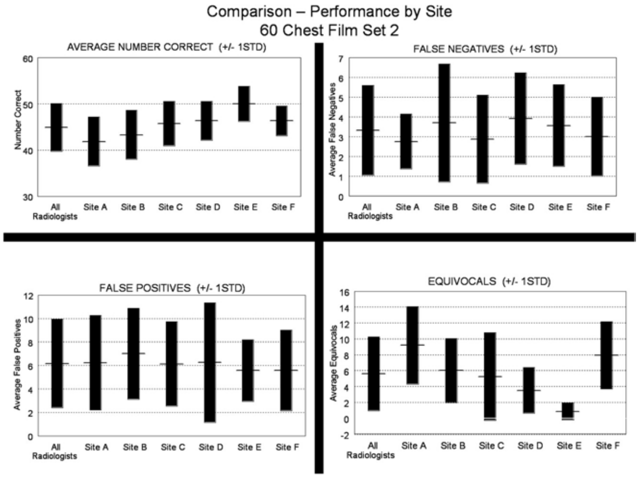

Fig 8. Comparison of different radiology groups by site or practice. Practices were in diverse states across the United States and England. The variation in performance was quite similar, with the exception of the number of equivocal interpretations at site E (see text). STD = standard deviation.

There was more conformity in the equivocation rate in group E, as demonstrated by the error bars. The amount of equivocation contributed to a lowering of the percentage of correct films. The group at site E suggested that the reason for their consistently low equivocation rate was that their local referring physicians had a low tolerance for equivocal radiology reports. Further research will be necessary to clarify the effect of referring physician culture on the performance of observers.

RESIDENTS' PERFORMANCE

My group has studied radiology resident performance for a number of years. In our studies, the greatest change in interpreting chest x-rays occurs between the first and second years of residency. After this, their performance as a group becomes somewhat stable. The change that occurs in training is primarily in the diminution of false-positive observations. Most senior residents had less equivocation and marked decreases in false-positive studies over the course of training. This serves to point out that the more novice observers and nonradiologists had considerably greater false-positive observations than the more experienced observers. Experience seems to increase confidence and accuracy. Figure 9 illustrates what happens during the course of training to be a radiologist. In this case, 3 groups of 10 to 14 radiology residents in 3 different years of training reviewed the set of 60 chest x-rays. The number of additional studies required and the number of false-negatives were the dominant changes seen in the performance of these observers. When comparing untrained radiology residents in their first year

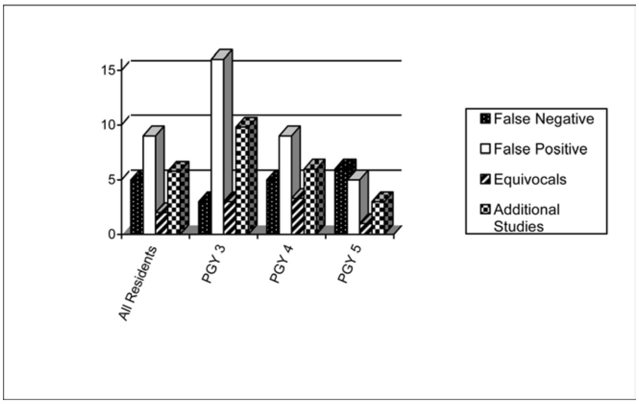

Fig 9. Comparison of resident performance on 60 chest x-rays, plain-film images (see text). PGY = physician graduate year.

with primary care physicians, a similar pattern of performance was observed. False-positives and additional studies were higher for nonradiologists and first-year residents than for other radiologists with more training.

CHARACTERISTICS OF EXTREME PERFORMERS

The shapes of the ROC curves vary with the characteristic decision making of the various groups of performers. My group compared the shapes of the ROC curves of high-performing and low-performing observers (Figure 10). When we compared the best-performing and the worst-performing radiologists, we noted improvement in all parameters, with the exception of the false-negative rate. False-negative rates do not distinguish the best performer from the worst performer in our experience. However, there was a substantial change in false-positive rates (3.2% vs 48.3%) between the best and worst observers. It is interesting to note that the average time per film observed was less for the top radiologist than for the worst-performing radiologist (0.69 vs 1.36 minutes). Reading these films faster did not equate with poorer performance. A major change seen between the best and worst performers was the number of additional films required to make definitive decisions (3.2% vs 31.7%). The inability to make a definitive decision was a significant problem for some observers when they studied this set of chest x-rays.

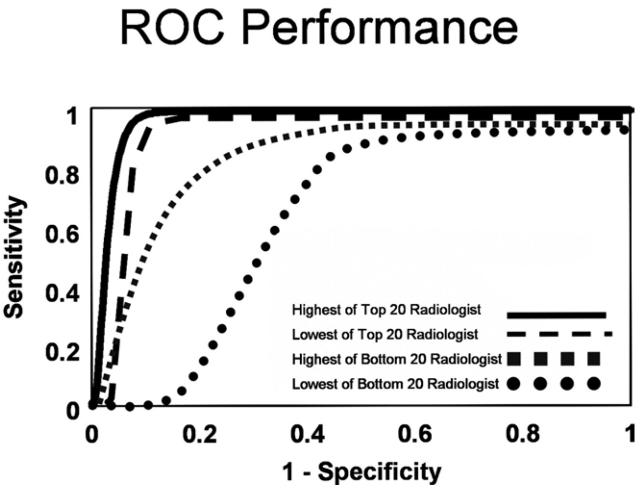

Fig 10. Range of performance between the top group of radiologists and the bottom group of radiologists. Note that the top 20 were quite similar, whereas the bottom 20 were quite dispersed. The variance in performance increased as performance deteriorated. ROC = receiver operating characteristic.

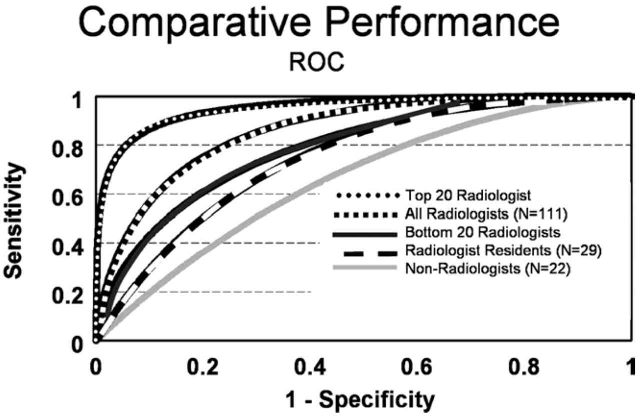

Fig 11. Comparative performance among board-certified radiologists, radiology residents, and nonradiologists. The performance of nonradiologists was much more random than that of trained radiologists. The worst radiologist was considerably better than the nonradiologists in terms of the area under the receiver operating characteristic (ROC) curve.

RADIOLOGISTS COMPARED WITH OTHER PHYSICIANS OBSERVING CHEST FILMS

My group compared the performance of 111 board-certified radiologists with that of 29 radiology residents and 22 physicians who were not radiologists (Figure 11). We have clearly demonstrated the difference between nonradiologists and radiologists. Nonradiologists performed worse than the lowest group of radiologists or radiology residents. However, the lowest group of radiologists performed worse than the average radiology resident.

CONCLUSION

Information can be defined as decreased uncertainty. This can be measured by observing a change in randomness. In a set of images, this change can be determined using ROC curves. The value added by any observer when interpreting a set of images can be measured by calculating the observer's ability to correctly divide the images into normal and abnormal sets. This calculation is represented by the area under an ROC curve.

Using a standard set of images, the performance of individuals and groups of individuals can be compared. The best performance can be made evident and can be used as a benchmark of an optimal standard to allow other performers to see what is possible. Some individuals, when made aware of how their decision making compares with that of others in their own groups, can improve their performance. By comparing groups of individuals, my group has been able to show how one group differs from another in performing the same task. Finally, this may allow for a determination of the value added to image interpretation by the process of becoming a radiologist. Further studies with larger groups will be necessary to confirm that impression.

REFERENCES

1. Potchen EJ. Prospects for progress in diagnostic imaging. J Intern Med 2000;247:411-24.
2. Zheng B, Chakraborty DP, Rockette HE, Maitz GS, Gur D. A comparison of two data analyses from two observer performance studies using Jackknife ROC and JAFROC. Med Phys 2005;32:1031-4.
3. Barnhart HX, Song J, Haber MJ. Assessing intra, inter and total agreement with replicated readings. Stat Med 2005;24:1371-84.
4. Monnier-Cholley L, Carrat F, Cholley BP, Tubiana JM, Arrive L. Detection of lung cancer on radiographs: receiver operating characteristic analyses of radiologists', pulmonologists', and anesthesiologists' performance. Radiology 2004;233:799-805.
5. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis and validation. Med Phys 2004;31:2313-30.
6. Marten K, Seyfarth T, Auer F, et al. Computer-assisted detection of pulmonary nodules: performance evaluation of an expert knowledge-based detection system in consensus reading with experienced and inexperienced chest radiologists. Eur Radiol 2004;14:1930-8.
7. Fuhrman CR, Britton CA, Bender T, et al. Observer performance studies: detection of single versus multiple abnormalities of the chest. Am J Radiol 2002;179:1551-3.
8. Brealey S, Scally AJ. Bias in plain film reading performance studies. Br J Radiol 2001;74:307-16.
9. Brealey S. Measuring the effects of image interpretation: an evaluative framework. Clin Radiol 2001;56:341-7.
10. Robinson PJ, Wilson D, Coral A, Murphy A, Verow P. Variation between experienced observers in the interpretation of accident and emergency radiographs. Br J Radiol 1999;72:323-30.
11. Krupinski EA, Evanoff M, Ovitt T, Standen JR, Chu TX, Johnson J. Influence of image processing on chest radiograph interpretation and decision changes. Acad Radiol 1998;5:79-85.
12. Giger M, MacMahon H. Image processing and computer-aided diagnosis. Radiol Clin North Am 1996;34:565-96.
13. Rossmann K. An approach to image quality evaluation, using observer performance studies. Radiology 1974;113:541-4.
14. Kundel HL, Lynch PR, Peoples L, Stauffer HM. Evaluation of observer performance using televised stereofluoroscopy. Invest Radiol 1967;2:200-7.
15. Shah SK, McNitt-Gray MF, DeZoysa KR, et al. Solitary pulmonary nodule diagnosis on CT: results of an observer study. Acad Radiol 2005; 12:496-501.
16. Potchen EJ, Cooper TG, Sierra AE, et al. Measuring performance in chest radiography. Radiology 2000;217:456-9.
17. Potchen EJ, Sierra AE. Value judgments in diagnostic radiology, how do we decide what to do? Radiology 1981;138:501-4.
18. Shannon CE. A mathematical theory of communication. Bell Syst Tech J 1948;27.
19. Towers JD, Holbert JM, Britton CA, Costello P, Sciulli R, Gur D. Multipoint rank-order study methodology: observer issues. Invest Radiol 2000;35:125-30.
20. Dorfman DD, Berbaum KS. A contaminated binormal model for ROC data: part III. Initial evaluation with detection ROC data. Acad Radiol 2000;7:438-47.
21. Rockette HE, Gur D, Metz CE. The use of continuous and discrete confidence judgments in receiver operating characteristic studies of diagnostic imaging technique. Invest Radiol 1992;27:169-72.
22. Gur D, Rockette HE, Good WF, et al. Effect of observer instruction on ROC study of chest images. Invest Radiol 1990;25:230-4.
23. Gur D, Rockette HE, Armfield DR, et al. Prevalence effect in a laboratory environment. Radiology 2003;228:10-4.
24. Swensson RG, King JL, Gur D. A constrained formulation for the receiver operating characteristic (ROC) curve based on probability summation. Med Phys 2001;28:1597-609.
25. Rockette HE, King JL, Thaete FL, Fuhrman CR, Slifko RM, Gur D. Selection of subtle cases for observer-performance studies: the importance of knowing the true diagnosis. Acad Radiol 1998;5:86-92.