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 nonthoracic 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.

© 2006 American College of Radiology
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

Download English Version:

https://daneshyari.com/en/article/4232649

Download Persian Version:

https://daneshyari.com/article/4232649

Daneshyari.com