Methodology and Application of Prospective Reader Studies: A Self-Assessment Module

Elizabeth A. Krupinski, PhD
Department of Radiology Research
University of Arizona

William C. Black, MD
Department of Radiology
Dartmouth-Hitchcock Medical Center

Felix S. K. Chew, MD, MBA
Department of Radiology
University of Washington

AJR2008:190:S29-S34, American Roentgen Ray Society

Methodology and Application of Prospective Reader Studies: Self-Assessment Module

Abstract

The educational objectives of this self-assessment module are for the participant to read selected sources on prospective reader study methodologies and to self-assess and improve his or her knowledge of this subject.

Introduction

This self-assessment module on prospective reader study methodologies has an educational component and a self-assessment component. The education component consists of four required articles that the participant should read. The self-assessment component consists of eleven multiple-choice questions with solutions. All of these materials are available on the ARRS Web site (www.arrs.org).

Educational Objectives

1. To understand the meaning and relevance of sample size and statistical power when designing reader studies as it pertains to both readers and images.
2. Understand the different types of study designs that can be used to conduct prospective reader studies as well as other measures of performance such as workflow.
3. Recognize when a Receiver Operating Characteristic (ROC) study is appropriate and what types of research questions it can answer.
4. To understand the basic nature of the Receiver Operating Characteristic (ROC) technique: use of rating scales, the ROC curve, area under the ROC curve (Az or AUC), interpreting metrics of performance.
5. Differentiate between true and false, positive and negative decisions.
6. Distinguish between common metrics of reader performance including sensitivity and specificity, accuracy, and positive and negative predictive value.
7. Recognize the importance of critical design issues in reader studies: reader experience, training, time between trials, types of images and types/subtlety of lesions, and environmental/viewing conditions.

Required Reading

1. Berger WG, Erly WK, Krupinski EA, Standen JR, Stern RG. The solitary pulmonary nodule on chest radiography: can we really tell if the nodule is calcified? AJR 2001;176:201-204.
2. Eng J. Receiver Operating Characteristic Analysis: A Primer. Acad Radiol 2005;12:909-916.
3. Kundel HL. History of Research in Medical Image Perception. J Am Coll Radiol 2006;3:402-408.
4. Metz CE. Receiver Operating Characteristic Analysis: A Tool for the Quantitative Evaluation of Observer Performance and Imaging Systems. J Am Coll Radiol 2006;3:413-422.

Supplemental Reading

1. Alpert HR, Hillman BJ. Quality and variability in diagnostic radiology. J Am Coll Radiol 2004;1:127-132.
2. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989;24:234-245.
3. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283-298.
4. Patton DD. Introduction to clinical decision making. Semin Nucl Med 1978;8:273-282.
5. Swets JA. ROC analysis applied to the evaluation of medical imaging techniques. Invest Radiol 1979;2:109-121.
6. van Erkel AR, Pattynama PMT. Receiver operating characteristic (ROC) analysis: Basic principles and applications in radiology. Eur J Radiol 1998;27:88-94.

INSTRUCTIONS

1. Complete the required reading.
2. Visit www.arrs.org and go to the left-hand menu bar under Publications/Journals/SAM articles.
3. Using your member login, order the online SAM as directed.
4. Follow the online instructions for entering your responses to the self-assessment questions and complete the test by answering the questions online.

Methodology and Application of Prospective Reader Studies: Self-Assessment Module

QUESTIONS

Question 1
Receiver Operating Characteristic (ROC) analysis techniques are often applied to studies of medical decision-making, especially in studies evaluating the impact of imaging technology on diagnostic accuracy. ROC analysis is based on which underlying theoretical framework.

A. Information-Processing Theory
B. Signal Detection Theory
C. General Recognition Theory
D. Bayesian Decision Theory
E. Information Systems Theory

Question 2
In which of the following scenarios would Receiver Operating Characteristic (ROC) analysis be appropriate?

A. An investigator wishes to determine if a new MRI reconstruction method improves the speed with which diagnostic decisions are rendered.
B. An investigator wants to determine if radiology technicians are at greater risk of cancer due to their exposure to ionizing radiation compared to the general population.
C. An investigator wishes to determine if radiologists can accurately detect calcification within pulmonary nodules with chest radiography.
D. An investigator wishes to examine the impact of a new drug on the survival rates of patients.
E. An investigator wishes to examine whether radiologists are more comfortable using a mouse or trackball as an interface device for softcopy reading.

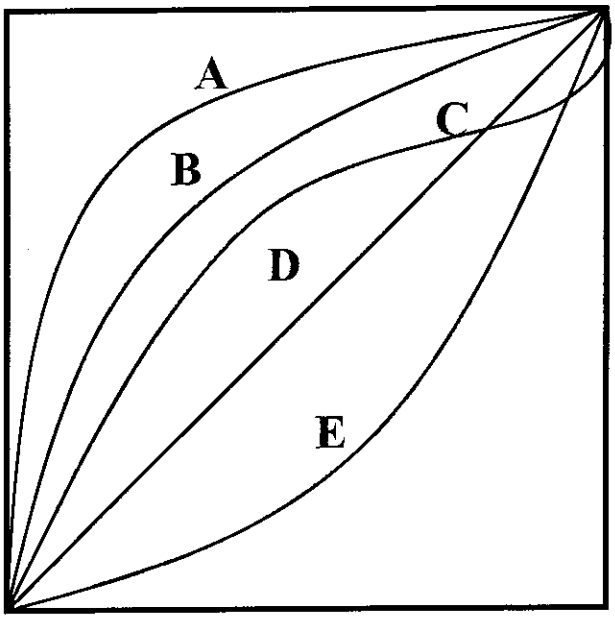

Figure 1. Set of ROC curves. The true positive fraction (sensitivity) is plotted along the vertical axis, and the false positive fraction (1-sensitivity) is plotted along the horizontal axis.

Question 3
Referring to Figure 1, which of the ROC curves (labeled a-e) represents an "improper" ROC curve or one that should be recalculated using "proper" methods?

A. A
B. B
C. C
D. D
E. E

Question 4
Referring to Figure 1, which of the ROC curves represents the best performance?

A. A
B. B
C. C
D. D
E. E

Question 5
One of the simplest metrics of diagnostic performance is accuracy (fraction of cases that are reported correctly). What is the main problem with using accuracy to report diagnostic performance?

A. Disease prevalence affects accuracy.
B. Sensitivity and specificity are not mathematically related to accuracy.
C. Calculating accuracy requires the use of calculus.

D. You cannot combine accuracy for multiple readers to get an overall measure of accuracy.
E. The number of cases affects accuracy.

Question 6
When choosing cases for an evaluation of observer performance, investigators typically select cases with only a single lesion. What is one main reason why this is done?

A. Using more than one abnormality per patient would be unrealistic.
B. Investigators are trying to avoid the Satisfaction of Search phenomenon.
C. ROC analysis cannot be used when there is more than one lesion.
D. Having more than one abnormality would increase the false positive rate.
E. Less experienced observers cannot be expected to find multiple lesions.

Question 7
When collecting a set of cases to include in a study of observer performance, why is it important to consider the number of cases to include?

A. It is necessary to replicate the typical reading volume a radiologist experiences
B. It is necessary to replicate the disease prevalence.
C. If you select too many cases, observers may become fatigued, resulting in underestimation of their true performance
D. Adequate sample size is necessary to achieve adequate statistical power.
E. Having too few cases does not give observers enough time understand the experimental task.

Question 8
When collecting a set of cases for an observer performance evaluation study, what is the best method to establish truth (or the reference standard) about whether a case contains an abnormality or not?

A. Ask a senior person to read all the cases and use those responses as truth.
B. Use the original official report for the case as truth.
C. Ask someone outside your institution to read the cases and use that as truth.
D. Let the lead investigator decide what truth is for each case since he or she is responsible for the study.
E. Use clinical or pathological data and follow-up results to establish truth.

Question 9
An investigator wants to add one of two new image-processing methods to a PACS workstation. An earlier ROC study showed both to yield equivalent observer performance. What type of study could be done to demonstrate that one method versus the other impacts reading efficiency?

A. Ask a set of clinicians which processing technique makes a set of images look better.
B. Process the same image with both techniques, display side-by-side and ask a set of clinicians to report which technique they would prefer to use.

C. Let the engineer who developed the techniques decide which one will be better in clinical use.
D. Have a set of clinicians interpret images processed with both techniques and time how long it takes for them to render a decision.
E. Put them both on the PACS workstation and record which technique is used more often.

Question 10
Readers of medical images can often be characterized as either conservative or liberal in their interpretations or use of decision thresholds. Which set of results characterizes a more conservative reader?

A. High true positives and high false positives
B. Low true positives and low false positives
C. High true positives and low false positives
D. Low true positives and high false positives
E. Equal amounts of true and false positives

Question 11
The National Council on Radiation Protection and Measurements scientific council under the guidance of Fryback & Thornbury developed a six-level hierarchical model for diagnostic efficacy. What are the six levels?

A. Technical, diagnostic, diagnostic thinking, therapeutic, patient outcome and societal
B. Technical, diagnostic, workflow, therapeutic, patient outcome and societal
C. Mechanical, diagnostic, workflow, therapeutic, patient outcome and societal
D. Diagnostic, therapeutic, curative, psychological, individual and societal
E. Diagnostic, therapeutic, curative, psychological, societal and global

SOLUTIONS

Solution to Question 1

The correct answer is b. Signal Detection Theory

All of these theories deal with the way that humans process information and render decisions. Signal Detection Theory deals specifically with the principle of signal detection (i.e., detecting a signal or target (e.g., tumor) in a background of noise (e.g., chest anatomy)) from a perceptual and decision-making perspective [1]. Information Processing Theory deals with the way the stages by which the human encodes, stores and retrieves information. It specifies four types of knowledge: general vs specific (useful in many tasks or specific ones), declarative (facts), procedural (how to), and conditional (when and how). General Recognition Theory is a perceptually based theory that attempts to explain how people identify objects, make decisions regarding their similarity, and make preference judgments. It has more to do with decision classification processes that occur after a target (lesion) has been detected. Bayesian Decision Theory is a specific type of Decision Theory that uses some prior distribution of data in the computation of the present decision, often taking utilities into account. Information Systems Theory deals with the transmission of information and ways to reduce the inherent uncertainty in information. It proposes a three-stage process to eliminate uncertainty through enactment (do something with the information), selection (decide which information to keep and which to ignore) and retention (what needs to be remembered).

Solution to Question 2

The correct answer is c. An investigator wishes to determine if radiologists can accurately detect calcification within pulmonary nodules with chest radiography. ROC analysis is used to assess decision accuracy in diagnostic detection tasks [2]. In a typical ROC study, observers are presented with a set of images, half containing a target of interest (e.g., a tumor) and half without any target (e.g., normal chest image). Ideally the set should contain a mix of subtle to moderately subtle targets. The images are randomized and each observer views the set in a given condition being tested (e.g., softcopy display with and without edge enhancement applied). The observer is asked to search each image and report whether a lesion/target is present or absent. They then report their confidence in that decision using either a discrete (5 or 6-point) or continuous (0 – 100) scale. The confidence ratings are then used to generate the ROC curve and the area under the curve can be calculated using standard methods. To measure speed (a) the investigator would use a stopwatch and compare times with a t-test. To compare risk or relative risk (b) of a group of people compared to the general population, one would use Relative Risk Analysis techniques. To study the impact of a drug (d) on patient survival, one would use Survival Analysis techniques. To examine tool use such as mouse vs trackball, one would rely on human factors analysis techniques such as a Time-Motion study.

Solution to Question 3

The correct answer is c. The ROC curve is not supposed to dip below the chance line (D) as curve C does. The chance line defines an observer who is essentially guessing about the status of an image and is therefore operating at "chance" (would call half the images normal and half abnormal). The area under the chance line by definition is 0.50. Ideally,

an observer who is qualified to carry out the experimental task and who understood the reporting instructions should perform better than chance and thus should have a resulting area under the curve higher than 0.50 (where 1.0 is perfect performance). "Proper" ROC methods have been developed to prevent this from occurring [3]. Curves A and B do not cross the chance line. Curve E falls below the chance line, typically indicating that the observer was not following the reporting instructions correctly (i.e., if the reporting scale used 1 = present, definite confidence and 6 = absent, definite confidence but the reader used 1 = absent, definite confidence and 6 = present, definite confidence).

Solution to Question 4

The correct answer is a. The closer an ROC curve is to the upper left corner the better performance is as generally measured by the Area Under the ROC Curve (Az) [4]. In ROC space, the area above the chance line (D) = 0.50 (The chance line defines an observer who is essentially guessing about the status of an image and is therefore operating at "chance" (would call half the images normal and half abnormal). The area under the chance line by definition is 0.50. Ideally, an observer who is qualified to carry out the experimental task and who understood the reporting instructions should perform better than chance and thus should have a resulting area under the curve higher than 0.50 (where 1.0 is perfect performance). Perfect performance (Az = 1.0) would be indicated by an ROC curve that follows precisely the left and upper lines. Curve B represents performance intermediate between A and C. Curve E represents performance below chance, typically indicating that the observer was not following the reporting instructions correctly.

Solution to Question 5

The correct answer is a. If a disease is relatively rare, occurring in only 5% of patients, then a clinician who calls all the cases negative will have an accuracy of 95% and that is misleading [4]. Answer b is incorrect because you can derive sensitivity and specificity from accuracy. Answer c is incorrect because accuracy is simply reported as a percentage and this does not require calculus. Answer d is incorrect because the number of cases in a test set by itself does not affect accuracy (although using too many cases in one setting could lead to reader fatigue and that could decrease accuracy). Knowledge of disease prevalence (or prior probability) in general can affect the decision criteria of a clinician. For example, coccidioidomycosis (Valley Fever) is caused by an organism found in the soil in the southwestern United States and affects primarily the lungs (nodules are observed). A clinician in New England who sees a patient with non-specific nodules is unlikely to consider coccidioidomycosis and thus misread the case of a patient who did not inform the clinician that they just returned from a vacation in Arizona, while a clinician in Arizona seeing the same nodular manifestations would be more likely to correctly consider coccidioidomycosis as the diagnosis.

Solution to Question 6

The correct answer is b. In Satisfaction of Search (SOS) observers do not report additional findings on images when they have found something suggested by the original search task [5]. For example, if the main task is searching for nodules in chest images, the presence of a rib fracture often goes unreported once the nodule is detected. Answer a is

incorrect since many patients do have multiple lesions per exam. Answer c is wrong because there are ROC techniques (e.g., FROC (Free-Response ROC), AFROC (Alternative Free-Response ROC)) designed specifically to account for multiple lesions. More than one lesion does not tend to increase the false positive rate, so d is incorrect. Answer e is incorrect because residents are trained from the beginning to search for and detect multiple lesions per case.

Solution to Question 7
The correct answer is d. Statistical power is the probability that one can reject the null hypothesis (there is no difference between conditions being compared) when it is indeed false (there is a true difference between conditions). Typically one wants a power of about 0.80, meaning that the probability that one can reject the null hypothesis as a result of the study is 80% [6]. Statistical power is affected significantly by sample size – the greater the sample size, the more power one typically has. Replicating reading volume (answer a) is impractical as is trying to replicate (answer b) prevalence (e.g., in screening mammography one may have 1 abnormal case per 1000). Although using too many cases may tire the readers and impair their performance (answer c), this is not the main reason to consider the sample size; rest breaks can always be incorporated into the protocol to address reader fatigue. Having written, clear instructions should avoid any observer confusion (answer e) regarding the task no matter how many cases are used.

Solution to Question 8
The correct answer is e. An independent assessment of truth using other types of clinical data is always preferred when these sources of data are available [7]. All of the other choices rely on a single observer to decide truth and that observer may be biased or simply incorrect. If other clinical data is not available to serve as the reference standard, the next best option is typically a panel of experienced clinicians.

Solution to Question 9
The correct answer is d. All of the other methods rely on the subjective opinion of individual observers about perceived image appearance and do not assess reading efficiency. Only method d actually records an objective measure of reader efficiency – time to render a decision [8].

Solution to Question 10
The correct answer is b. A conservative reader typically adopts a relatively high decision threshold (must have a lot of evidence to report an abnormality as present), resulting in fewer positive decisions (both true and false) than a more liberal reader [9]. A liberal reader has a lower decision threshold and thus is characterized by answer a. Since the decision threshold affects both true and false positive decisions the same way (when one goes up the other does too), answers c, d and e rarely occur in experimental settings.

Solution to Question 11
The correct answer is a. A diagnostic test is considered technically effective if its result is accurate and precise in a physical sense [10]. Diagnostic efficacy concerns the extent to which the results of a diagnostic test agree with patients' actual states of health.

Diagnostic-thinking efficacy is difficult to measure but is the extent to which a diagnostic test affects physicians' subjective estimates of disease likelihood. Therapeutic efficacy addresses the question of how and by how much does a particular diagnostic test changes the way in which patient are treated. Patient-outcome efficacy refers to whether a patient's health is demonstrably improved by use of the test. Societal efficacy merges private and public considerations (e.g., cost/benefit/effectiveness) to assess diagnostic tests within the context of the social endeavor. Mechanical, workflow, psychological curative and global efficacy are not part of the framework. This framework is valuable in today's clinical environment because it acknowledges that it is no longer sufficient to simply demonstrate that a new technology can better depict anatomy, function, disease etc. and thereby improve diagnostic accuracy. The decision whether to adopt or forego a new technology also depends on its cost, not only in the monetary sense but also in the societal sense and the outcomes effected by the new technology.

REFERENCES

1. Swets JA, Pickett RM. Evaluation of Diagnostic Systems: Methods from Signal Detection Theory. New York, NY: Academic Press; 1982.
2. Berger WG, Erly WK, Krupinski EA, Standen JR, Stern RG. The solitary pulmonary nodule on chest radiography: can we really tell if the nodule is calcified? AJR 2001;176:201-204.
3. Pan X, Metz CE. The "proper" binormal model: parametric receiver operating characteristric curve estimation with degenerate data. Acad Radiol 1997;4:380-389.
4. Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978;8:283-298.
5. Kundel HL. History of Research in Medical Image Perception. J Am Coll Radiol 2006;3:402-408.
6. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol 1989;24:234-245.
7. Revesz G, Kundel HL, Bonitatibus M. The effect of verification on the assessment of imaging techniques. Invest Radiol 1983;2:194-198.
8. Krupinski EA, Roehrig H, Dallas W, Fan J. Differential use of image enhancement techniques by experienced and inexperienced observers. J Digit Imaging 2005;18:311-315.
9. Patton DD. Introduction to clinical decision making. Semin Nucl Med 1978;8:273-282.
10. Metz CE. Receiver Operating Characteristic Analysis: A Tool for the Quantitative Evaluation of Observer Performance and Imaging Systems. J Am Coll Radiol 2006;3:413-422.