**Communications of the Association for Information Systems**

Volume 16 Article 16

August 2005

# Information Systems and Health Care IV: Real-Time ROC Analysis to Evaluate Radiologists' Performanceof Interpreting Mammograpny

Min Wu
*University of Wisconsin at Milwaukee*, wu@uwm.edu

Etta Pisano
*University of North Caroline at Chapel Hill*, etpisano@med.unc.edu

Yuanshui Zheng
*University of North Caroline at Chapel Hill*, yzheng@med.unc.edu

Communications of the Association for Information Systems

# INFORMATION SYSTEMS AND HEALTH CARE – IV: REAL-TIME ROC ANALYSIS TO EVALUATE RADIOLOGISTS' PERFORMANCE OF INTERPRETING MAMMOGRAPHY

Min Wu
Department of Health Sciences
University of Wisconsin at Milwaukee
wu@uwm.edu

Etta Pisano
Yuanshui Zheng
Department of Radiology
University of North Carolina at Chapel Hill

**ABSTRACT**

This paper describes how to use Receiver Operator Characteristic (ROC) analysis to evaluate radiologists' performance of interpreting digital mammograms in real-time. We developed an experimental testing system, which implemented a set of clinical lesion-matching rules to prepare raw ROC data. The system can automatically provide detailed evaluations of the performance, such as sensitivity, specificity, positive predictive value, negative predictive value, diagnostic accuracy, ROC curve, and area under the curve (Az). Based on a preliminary evaluation of the system, we found that ROC analysis is appropriate for a real-time computer application, directly using the raw data from a database, to evaluate the performance of radiology residents.

**KEYWORDS:** ROC, evaluation, medical education, mammography

## I. INTRODUCTION

Breast cancer is second to lung cancer in the cancer mortality rate among U.S. women today. Early detection is the best defense, and mammography is the best tool to detect early stage breast cancer. To improve the quality of care, increasingly strict auditing of radiologists' performance is being proposed. For example, the US Congress is discussing whether radiologists interpreting mammograms should be required to do self-evaluation of the interpreting skills in mammography as part of their Continuous Medical Education (CME) in the near future.

Diagnostic performance of radiologists is normally measured by sensitivity and specificity:

- Sensitivity is the probability of detecting a disease when a disease exists. High sensitivity is important to achieve the maximal detection of diseases at an early stage, where treatment offers an increased potential for cure.

- Specificity is the probability of identifying negative images correctly when a disease does not exist. High specificity is also desirable to avoid unnecessary biopsies, with their associated financial and emotional costs to patients.

The Receiver Operator Characteristic (ROC) curve analysis demonstrates trade-off between sensitivity and specificity graphically. The intrinsic meaning of the area under the ROC curve is: the averaged sensitivity over all specificity.

ROC methods were first applied rigorously to medical imaging [Lusted, 1971]. Currently, the ROC analysis method is a well-known method for radiologists [Metz, 1986]. The ROC applications in Radiology involve the readers (radiologists), imaging machines, and images. In previous projects [Swets, 1979; Huber et al., 1998; Metz, 1999a; Armato et al. 2003 ; Shusuke et al. 2003], the ROC methods were mainly applied to evaluate the imaging machines or compare the image post-processing algorithms, using the same readers (radiologists). For example, ROC methods were applied to evaluate computer algorithms in computer aided diagnoses applications [MacMahon et al., 1999; Abe et al., 2003]. Computers could provide suggestions for radiologists to interpret the images, using various artificial intelligence approaches, such as neural networks.

Recent ROC research focuses on evaluating radiologists using fixed-image cases. However, there are many challenges in using ROC analysis for evaluating radiologists' performance in real-time. For example, the ROC method requires truth on all cases, including negatives. A complete determination of the "truth" about negative cases, interpreted by radiologists in clinics, requires follow-up data of at least 1-year. Operational issues of the ROC analysis method need to be explored; for example, how to integrate the work-flows of the clinical practices into a computer-based system.

The National Library of Medicine (NLM) funded the National Digital Mammography Archive (NDMA) project since 1998. A tele-educational system developed at University of North Carolina at Chapel Hill (UNC-CH) was an important component of the NDMA project [Wu et al., 2002]. The system includes an annotation tool, a case demonstration tool, and a testing tool for radiology residents and breast-imaging fellows. The testing tool uses the Receiver Operating Characteristic (ROC) curve methodology to evaluate radiology residents' performance of interpreting mammograms.

It is really a challenge to create ROC curve results automatically, based on the raw data in the database in a short time. However, if the system is successfully developed in this setting, we would propose their use to the American Board of Radiology for certification of radiologists for professional competence in Breast Imaging. This program could serve as a model for how softcopy breast imaging competency examinations might be conducted in the future.

This paper proposes an experimental system design (Section II), describes a preliminary evaluation of the system (Section III), and discusses the results of the study (Section IV).

## II. EXPERIMENTAL SYSTEM

The goal of the testing tool is to evaluate two different skills of radiologists in the diagnoses of breast cancer.

1. Detecting abnormalities visible in mammograms.
2. Diagnosing the detected abnormalities.

Before the testing tool was used by radiology students, mammogram cases were systematically collected, validated, and stored into an education database [Wu et al., 2002]. Radiology faculty used an annotation tool to annotate lesions in the mammogram cases graphically based on pathological reports [Zheng et al., 2004]. A work-flow of the testing tool was carefully designed for ROC analysis (Figure 1). The processes in this real-time ROC application are discussed in detail in the following subsections.

## DETECTION OF ABNORMALITIES

The first task of radiology residents using the testing tool is to detect abnormalities in mammograms. In traditional studies of breast imaging, the location of each lesion (abnormality) can be coded into "side", "AP location" (Anterior, Central, Posterior) and "O'Clock location". The verbal description of the location of a 3-D point on the 2-D images, which are just optimal guesses by radiologists using the "O'Clock location", frequently cause errors. In addition, more errors can be introduced when relatively complex rules are implemented to match lesion diagnosis to pathologic data, because the location-method for lesion localization utilizes clock face or quadrants, which may not match the pathologist's and surgeon's perception of the location of a lesion in the breast. The graphical annotation of lesions in our computer-based education system successfully solved this location problem. Instead of verbal descriptions,

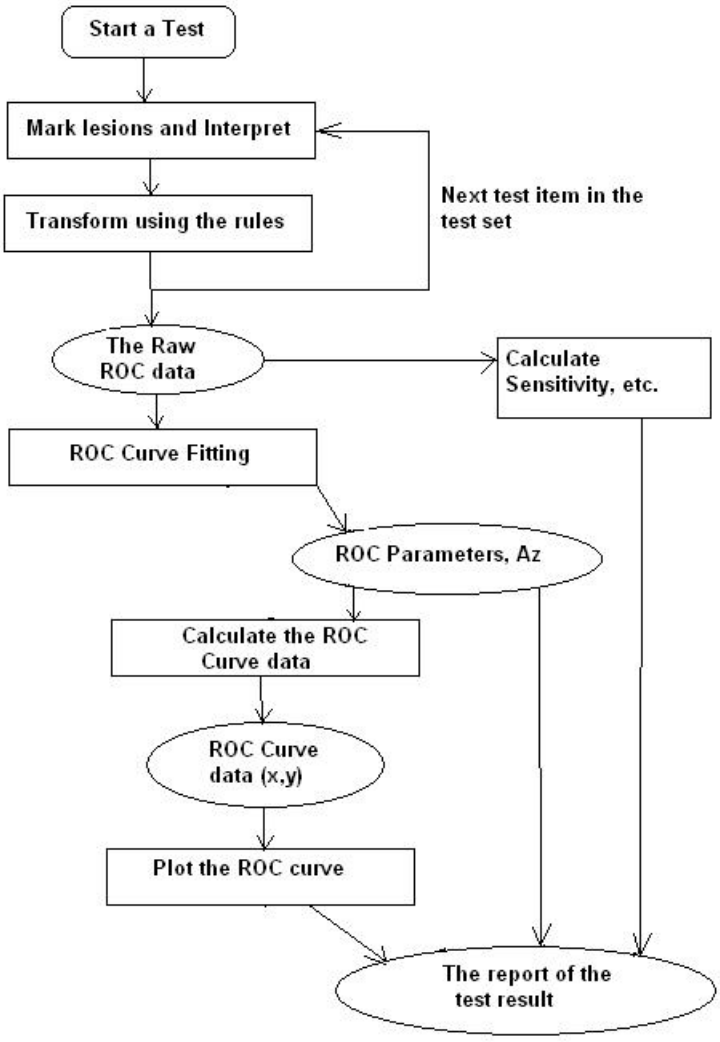


Figure 1. An Overview of Testing Tool Work-Flow

radiologists' interpretations and mammograms are all visualizations of woman's breast information. The locations of lesions that exist in breasts are projected onto a two-dimensional mammogram, preserving the relative positions and relative distances of rendered objects. The location of a lesion is coded as a spatial data set of coordinates x and y. The testing tool provides an interface for detecting the abnormalities in mammograms (Figure 2). The test taker can review the images, and adjust the contrast by using an intensity window at the bottom of the

screen. In this single-image mode, the test taker can mark a lesion by using the middle button of a mouse (Figure 3).

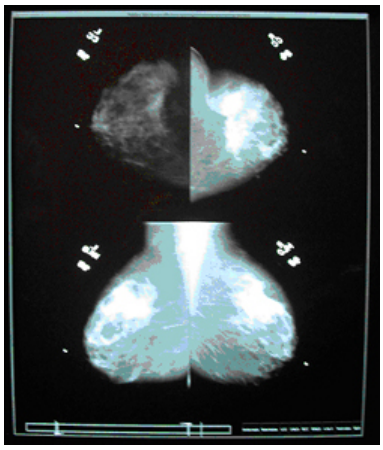

Figure 2. The Interface of the Testing Tool to Display Mammograms

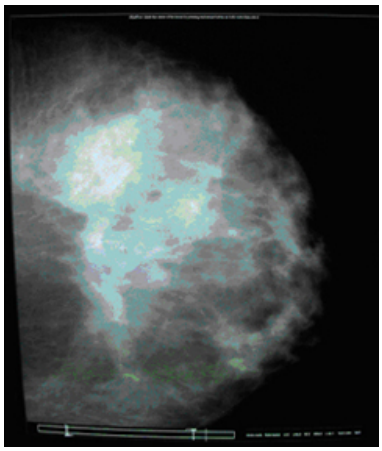

Figure 3. The Single Image Mode Interface on the Testing Tool

## DIAGNOSE THE DETECTED ABNORMALITIES

After detecting abnormalities in mammograms, the second task of radiology residents using the testing tool is to diagnose the abnormalities, and make decisions about the detected abnormalities. Specifically they must decide whether the lesion is benign or malignant. In a practical clinical setting, only a few discrete diagnostic categories are used for interpreting mammograms. For example, the American College of Radiology (ACR) designed the Breast Imaging Reporting And Diagnosis System (BI-RADS), which is a 5-category confidence-rating scale for interpreting mammograms in a clinical setting [ACR, 1998].

Our testing tool was designed to use ROC curve methodology to evaluate radiology residents' performance to interpret mammograms. ROC analysis is degraded by using a limited number of discrete categories as opposed to a continuous rating scale [Wagner et al., 2001]. The use of a limited number of discrete categories, such as BI-RADS 5 categories, may lead to a poor sampling of the large region of the ROC curve. Since radiologists use the confidence-rating scale in a conservative way, they do not distribute their responses more or less uniformly over all categories. The ROC curve may consist of vertical and horizontal line segments on a conventional plot. These segments are called "degenerate" data sets.

The use of a continuous, quasi-continuous confidence-rating scale for ROC data collection can reduce the likelihood of "degenerate" data sets [Rockette et al., 1992]. Therefore, we proposed a new confidence-rating scale in our ROC testing tool with a continuous slide bar. However, radiology residents were regularly trained to interpret abnormalities by using verbal descriptions, such as BIRADS. That is, they are familiar with reasoning using traditional verbal descriptions. Even through the new innovative slide bar is, in our opinion, intuitive and easy, it is still something new for physicians. Consequently, we improved the new continuous rating scale by combining it with a 7 category verbal description. At this stage, we provide both a 7-categories verbal scale and a continuous visual scale (Figure 4). A test taker interprets the lesion by moving the slider bar. A probability of malignancy for the lesion is recorded.

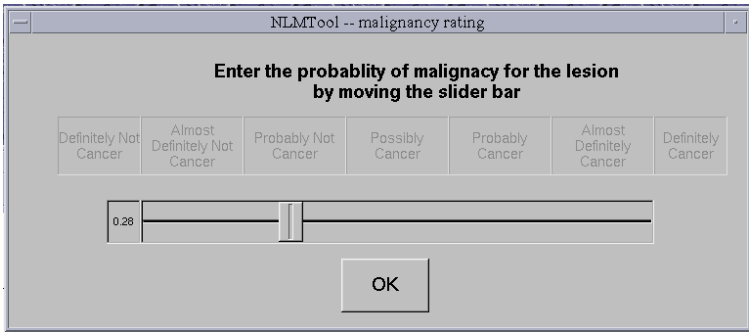


Figure 4. An Interface for the Test Taker to Rate a Lesion

## LESION-MATCHING RULES FOR ROC ANALYSIS

Before the testing, radiologist teachers used an annotation tool to annotate lesions graphically in the mammogram cases based on pathological reports. The "truth" of lesions was stored in our database. During the computer-based testing, test takers (radiology residents) can draw a mark on the digital mammogram to identify the abnormality. The test taker's goal is to place the mark within the circle of the lesion annotated by the faculty. The testing tool can directly compare the graphical annotations of the lesion between the faculty and test taker.

The clinical setting in a real world is always complex. Each mammogram case for a patient includes many images, because many 2-D images are taken from different views, such as vertical (CC) and horizontal (MLO). Each image may show more than one lesion. Because a computer application just does what humans tell it to do, the testing tool needs clearly defined lesion-matching rules to follow. Rules were proposed for the computer-based testing tool to prepare ROC raw data (Figure 5).

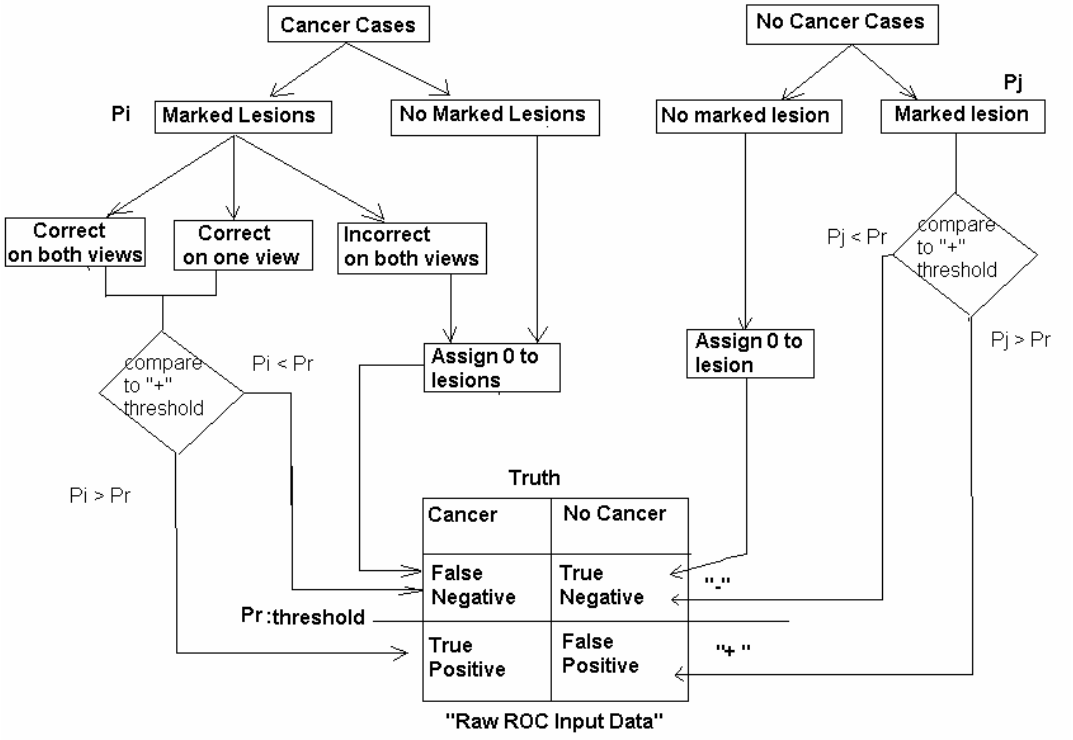


Figure 5 The Rules for Preparing ROC Raw Data

**Confidence Threshold (Pr)**

A confidence threshold (Pr), which separates "positive" decisions from "negative" decisions, should be defined at the very beginning. Then, the resulting sensitivity and specificity are calculated for that cutoff. In the experimental system, a probability of 0.5 was chosen as a threshold (cutoff point), corresponding to the "possible cancer". The 50% probability as the boundary between "positive" and "negative" decisions was chosen based on the middle category of BI-RADS for "possible cancer" and a probability of a pure guess, while there are fewer malignant cases in the usual clinical situations than in the tests. If the probability of malignancy of a lesion in the test-taker's interpretation is less than 0.5, it counts as a negative test. If the probability of malignancy of a lesion is larger than 0.5, it counts as a positive test.

**Rules for a "Cancer" Case**

The rules for a **"Cancer"** case scenario are defined as follows.

1. A reader marks a cancerous lesion in that case and provides an interpretation of probability (Pi). The reader will get credit for finding the cancer if he or she marks it correctly on only one view. After the system defines a confidence threshold **(Pr)** to separate the "positive" decisions and "negative" decisions, the tool compares Pi to Pr. If Pi is above the threshold (Pr), the Pi will be a "true positive" result. Otherwise, if Pi is below the threshold (Pr), it will be a "false negative" result.

2. If the reader misses the lesion on both views, the reader's interpretation (Pi) will be assigned to 0 for that lesion, as a "false negative" result.

3. If the reader does not mark anything and does not provide any interpretation about the lesion, the system will assign a probability of "0" for that lesion, as a "false negative" result.

**No Cancer Case Scenario**

The rules for a "No Cancer" case scenario were as follows.

1. A reader marks a lesion in that "No Cancer" case and provides an interpretation of probability (Pj). The system directly compares Pj to the confidence threshold (Pr). If Pj is above the threshold (Pr), the Pj will be a "false positive" result. Otherwise, if Pj is below the threshold (Pr), it will be a "true negative" result.

2. If the reader does not mark anything in this "No Cancer" case and does not provide any interpretation about the case, we will assign a probability of "0" for that case, as a "true negative" result.

*Results of ROC analysi***S**

A series of ROC analysis results are calculated by the testing tool based on the raw ROC data, such as, Sensitivity, Specificity, Positive Predictive Value (PPV), Negative Predictive Value (NPV), Diagnostic Accuracy and Az value (area under the curve).

- PPV is a measure of the percentage of cases correctly identified by the test taker as actually containing cancer.

- Negative Predictive Value (NPV) is a measure of the percentage of cases correctly identified by the test taker as NOT containing cancer.

- Diagnostic accuracy is a measure of the test taker's ability to arrive at a correct diagnosis for all cases in the test.

Specially, the Az value from ROC analysis is an appropriate index of the radiology residents' performance. A ROC curve-fitting software program, PROCROC[1], was integrated successfully into the test tool to generate the estimation of the area under the ROC curve (Az value) and other parameters about the curve. The errors of the estimated area under the ROC curve are largely reduced by using a continuous scale for data collection and the curve-fitting algorithm in PROCROC. The testing tool plots an ROC curve on the computer screen at the end of testing (Figure 6).

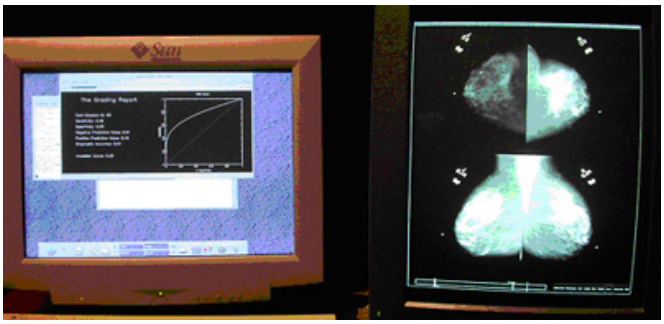

Figure 6. The Experimental ROC Testing Tool

## III. PRELIMINARY EVALUATION

A preliminary evaluation of the testing tool for radiology residents interpreting mammography was conducted at UNC. This evaluation is introduced and analyzed in this section.

### EVALUATION METHOD

Four persons were recruited to use the testing tool, and each person took two tests. The two tests were systematically constructed at UNC-CH. Each test consists of more than thirty cases in each test set [Wu et al., 2004]. The cases in each test are different, while the contents of the two tests are similar. The percentages of "normal (no findings)" cases, "benign follow-up" lesions, "benign biopsy" lesions and "malignant biopsy" lesions in each test are 20%, 20%, 30% and 30%. The four test takers for the evaluation studies came from different training backgrounds.

- Test taker A is a graduate student in biomedical engineering,
- Test taker B is a medical school student. Neither test taker A nor B had any mammogram training.
- Test taker C is a 1st year radiology resident, and
- Test Taker D is a radiologist in practice.

Both Test taker C and D were trained in breast imaging.

The open-test used allows test takers to bring their books or notes when taking a test. A magnifying lens was also provided to all test–takers. No time limit was imposed on the test takers.

Before the test, each test taker read a brochure about the testing tool, which includes information about test content specifications, testing goals, time requirement of the testing, the processes of the testing and an explanation of terminologies in the grading reports. Before taking the test, each user practiced through a sample test with several cases for test takers to familiarize

[1] PROCROC was developed by Drs. Pan and Metz at the University of Chicago [Metz et al., 1999b].

themselves with how to use the testing tool interfaces. After the test taker finished all cases in the test, a grading report was created automatically and displayed on the screen. Finally, the users of the testing tool were encouraged to fill out an online survey about the system.

As a result, all test takers successfully completed the two tests and received their grading reports immediately. Scores for Test II were lower than Test I. We found that the main factor explaining this difference is that the image qualities differed in the tests. Both untrained and trained participants completed tests within 2 hours or less. The results of the test takers' grading reports include sensitivity, specificity, PPV, NPV, diagnostic accuracy, and Az value, reported with 95% confidence intervals.

## EVALUATION RESULTS

The correlation coefficient between untrained participants A and B is 0.874 with significance at 0.01, based on the test results, and the correlation coefficient between trained participants C and D is 0.977 with significance at 0.01.

The sensitivity in most published mammography audits is greater than 85% in clinical practice [Rockette et al., 1992]. The system found that test taker A and B achieved almost the same sensitivities (0.30 vs. 0.31), finding 2 or 3 lesions from the 10 malignant lesions included in each test (Figure 7). Both test taker A and test taker B were not trained in breast imaging. Therefore, their scores represent a baseline of sensitivity in untrained population of test takers. In Test I, the sensitivity of test taker C (the 1st year resident) is 0.38, and that of test taker D (the radiologist) is 0.46. The sensitivity scores provided by the testing tool met our expectation that test taker C and D would find more abnormalities than test taker A and B. The results of test takers in Test II are similar.

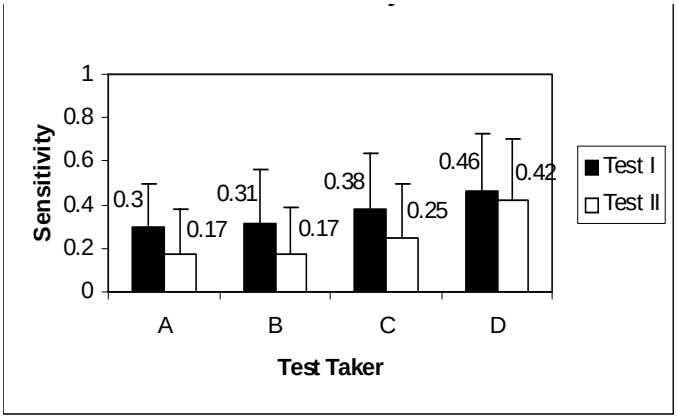


Figure 7. Sensitivities of Test Takers

The specificity is usually found to be greater than 90% in clinical practice [Spring et al., 1987; Bird, 1989; Sickles, 1992; Linver et al., 1992]. We found that test takers A and B found 5 or 6 lesions from 10 benign lesions (see Figure 8). The specificity of test taker C (the 1st year resident) is 0.81 in Test I, and the specificity of test taker D (the radiologist) is 0.78. That is, the resident and the radiologist may call 2 or 3 of the 10 benign lesions as malignant. The specificity scores, provided by the testing tool, demonstrated that the specificities of all test takers are very close between 0.62 and 0.81 in Test I. On average, test taker C and D did slightly better than test taker A and B.

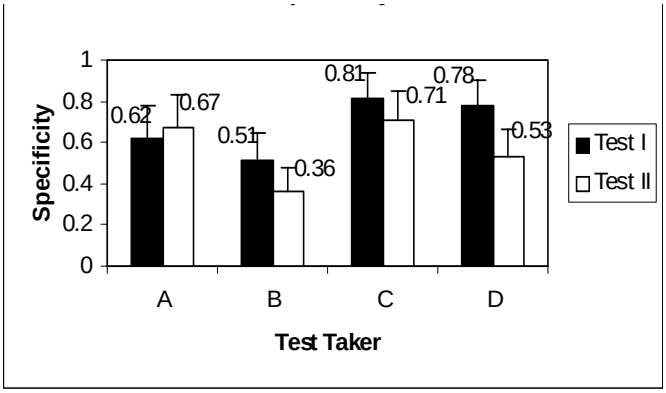


Figure 8. Specificities of Test Takers

A high PPV is desirable, since it is a measure of how likely a positive result is indeed a true positive. The overall PPV of first-screening mammography is reported to be 5% to 38% [Kopans, 1994; Kerlikowske et al., 1993]. The PPV is dependent on the radiologists' criteria for biopsy or follow-up. The proper positive predictive value for cancer in biopsies of nonpalpable lesions diagnosed on screening mammograms should be 30 to 40 percent, however, that will cause diagnosis of some cancers to be delayed [Hall et al., 1988]. We found that test taker A and B picked up 2 or 3 malignant lesions from 10 lesions with positive diagnoses (see Figure 9). In Test I, the PPV of test taker C (the 1st year resident) is 0.42, while PPV of test taker D (the radiologist) is 0.4. The PPV scores, provided by the testing tool, demonstrate that PPV of test taker C and D are better than those of test taker A and B.

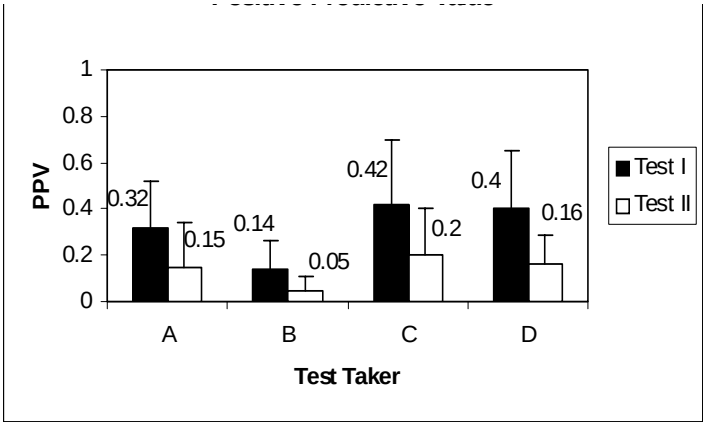


Figure 9. Positive Predictive Values of Test Takers

A high NPV is desirable, because it is a measure of how likely a negative result is indeed a true negative. Test taker A and B correctly identified 6 or 7 benign lesions from 10 lesions without cancer (see Figure 10). In Test I, the NPV of test taker C (the 1st year resident) is 0.78, while NPV of test taker D (the radiologist) is 0.82. The NPV scores, provided by the testing tool, demonstrated that NPV for all test takers are very close between 0.6 and 0.82 in Test I. On average, test taker C and D did slightly better than test taker A and B.

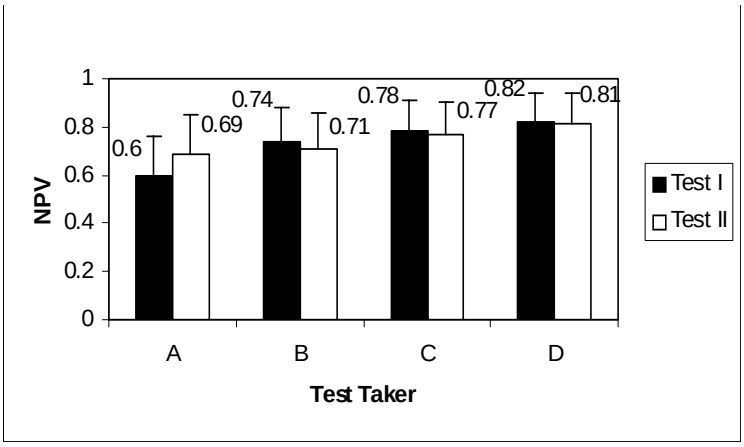


Figure 10. Negative Predictive Values of Test Takers

High diagnostic accuracy is desirable. We found that test taker A and B made approximately 1 correct diagnosis in every 2 diagnoses (see Figure 11). In Test I, the diagnostic accuracy of test taker C (the 1st year resident) is 0.69, while the diagnostic accuracy of test taker D (the radiologist) is 0.7. The diagnostic accuracy scores, provided by the testing tool, demonstrated that test takers C and D performed overall better than did test takers A and B.

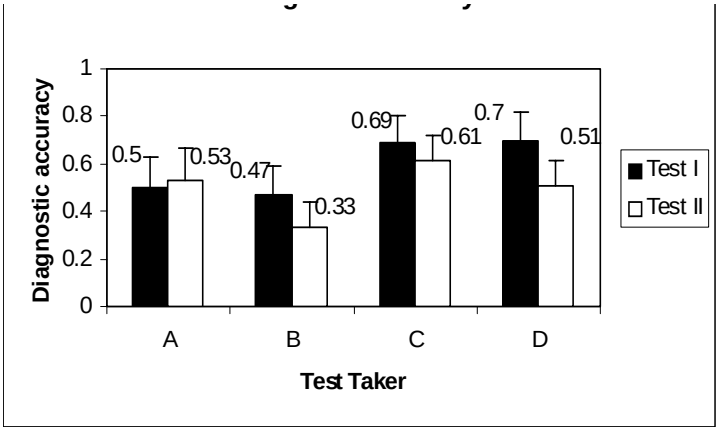


Figure 11. Diagnostic Accuracy of Test Takers

In our grading reports, ROC Curves and Az value are new indexes in terms of the performance of test takers. Our testing tool implemented ROC analysis methods to estimate the performance of test takers. When the confidence threshold is varied, an ROC curve is traced over the range of confidence thresholds.

Az value is the sensitivity averaged over all specificities. It is commonly used as a summary measure of diagnostic accuracy, which takes on values from 0 to 1. High Az value is desirable. Reported Az values for radiologists in screening mammography vary widely, with reports of Az values of 0.61 [Jiang et al., 1999], 0.76 [Lewin et al., 2001], 0.81 [Kacl et al., 1998], 0.83 [Lin et al., 1995], 0.85 [Taplin et al., 2000], 0.84-0.89 [Poon et al., 1992] and 0.94 [Sahiner et al., 1998]. The variability is likely due to differences in radiologist ability and variation in the degree of diagnostic difficulty of the databases used for testing [Nishikawa et al., 1994]. The tool showed that the Az values for test taker A and B were low, around 0.22 in Test I. The Az scores of test taker C and D are much larger than those of test taker A and B. The differences of Az values between Untrained group (0.12~0.30) and Trained group (0.64~0.86) are statistically significant (Figure 12). In addition, the testing tool demonstrated that the Az values are more sensitive than

other parameters to estimate test takers' skills in interpreting the mammograms. We believe that Az value will be an appropriate index of test taker's performance. Our testing tool successfully provided Az value for each test taker.

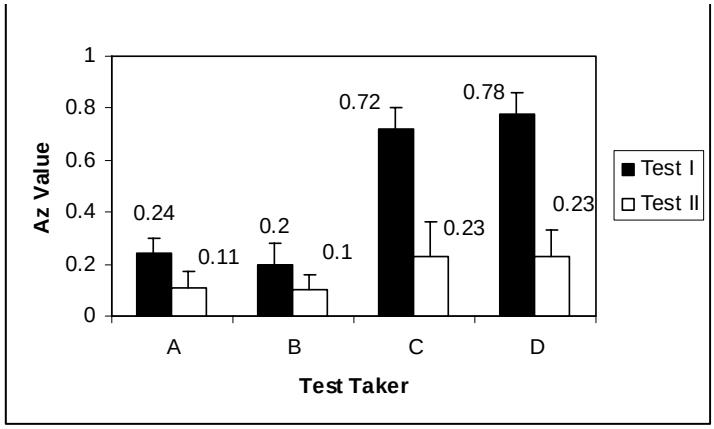


Figure 12. Area Under Curve of Test Takers

**The ROC Curve**

As mentioned previously, the ROC curve is a graphic method for showing the trade-off between the sensitivity and specificity of a test. A high specificity and relatively low sensitivity for a radiologist implies that he or she is more conservative in calling patients back in order to avoid unnecessary biopsies. A radiologist with high sensitivity and relatively low specificity, is more aggressive in calling patients back in order to achieve the maximal detection of cancers at an early stage. Also, shapes of ROC curves provide useful information about the radiologist's performance. In Test I, a ROC curve of test taker C is shown in Figure 13, and a ROC curve of test taker D is shown in Figure 14.

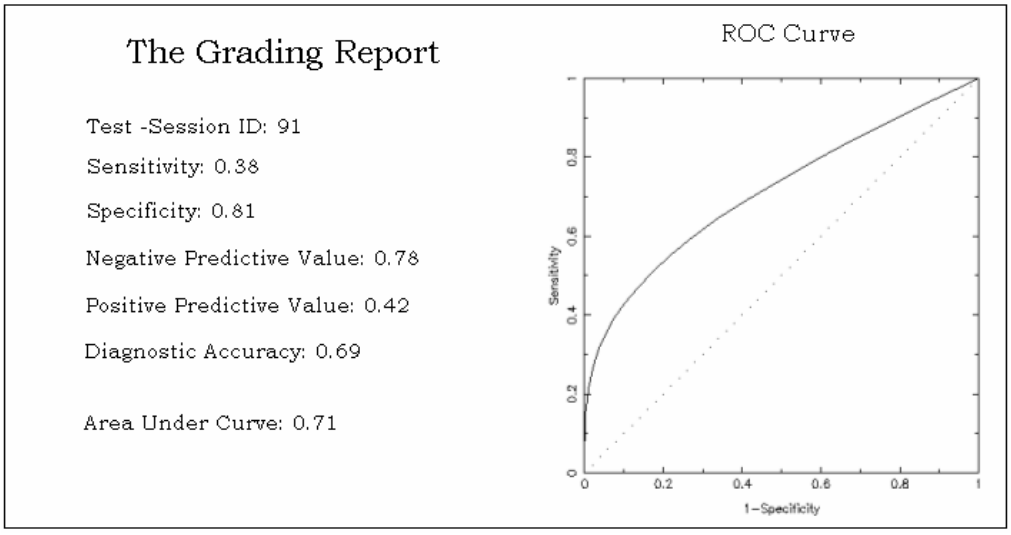


Figure 13. The ROC Curve of Test Taker C on Test I

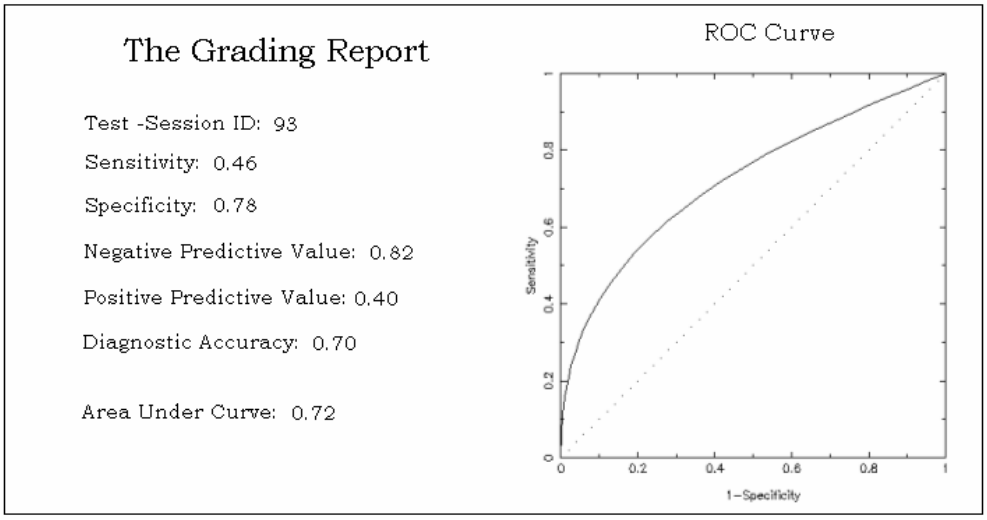


Figure 14. The ROC Curve of Test Taker D on Test I

Even though Az values in Figure 13 and 14 are almost the same, the shapes of ROC curves are slightly different. We can see that test taker C's ROC curve in Figure 13 is more convex to the left side of high specificities than test taker D's ROC Curve in Figure 14. An optimal threshold is associated with a slope. The system can calculate a slope, using an equation (1) [Zweig et al., 1993].

$$M = (FPC/FNC)*((1-p)/p) \quad (1)$$

FPC and FNC are the relative cost of False Positive (FP) and False Negative (FN) errors; and p is the prevalence of positive cases. If the ROC plot is a smooth and parametric curve, M describes the slope of a tangent to this curve. The point at which the tangent touches the curve identifies a particular sensitivity/specificity pair. Because the shapes differ between test taker C and D, different points will be found for the same slope value, suggesting that optimal thresholds of test taker C and D are different. If FPC > FNC, the threshold should favor specificity, while sensitivity should be favored if FNC > FPC.

## IV. CONCLUSIONS AND DISCUSSIONS

Currently, ROC analyses of evaluating radiologists' diagnostic performance are not real-time analyses. Instead these analyses are based on derived data, which are counted and collected manually from a clinical database. In this paper, the methodology issues of real-time computer-based ROC analysis for evaluating radiologists' performance are discussed and a prototype design of a real-time ROC testing tool is developed and evaluated. The real-time testing tool would systematically quantify and verify the quality of the diagnostic accuracy of radiology residents interpreting mammograms, and automatically provides test takers with detailed performance test results, such as sensitivity, specificity, positive predictive value, negative predictive value, diagnostic accuracy, ROC curve, and area under curve (Az value). We believe that the ROC analysis is appropriate for a real-time computer application, directly using the raw data from a database, to evaluate the performance of radiology residents.

Based on the preliminary evaluation, we observed the effectiveness of the test tool. The scores on the grading reports of test takers provided by the tool successfully showed the differences in the performance of test takers who interpreted mammograms. In particular, the Az values in ROC analysis clearly distinguish the "no-training" test takers from the "trained" ones.

The new system presents certain challenges to test takers, because there are more malignant cases in the tests than those detected in the usual clinical environment, and test takers are evaluated lesion by lesion for accuracy purpose, instead of case by case. Furthermore, computer test items and formats of our testing tool are somewhat new to test takers. Specifically, the radiology resident and radiologist had to translate their diagnosis from the clinical category rating scale with which they are very familiar with to our continuous rating scale. Within a short test time, test takers are required to detect all kinds of lesions and classify them. Previous research showed that quantities of cases interpreted (reader volume) affects cancer-detection accuracy [Esserman et al., 2002]. In our test setting, test takers are exposed to more than 30 cases in a single test within 2 hours, while the minimum annual reading volume in the United States is just 480 as set by the Mammography Quality Standards Act of 1992 [Food and Drug Administration, 1997]. The test scores of their overall performance provided by our system would help test takers to identify their strengths and weaknesses. Test takers in the preliminary evaluation expressed a willingness to review the literature on mammography again and the mammographic appearance of breast cancers.

In our test tool design, we employed ROC curve for evaluating test takers' comprehensive skills of interpret mammograms. To simplify the comparison we plot the curves in two dimensions using the usual axes for ROC curves, namely true and false positive conditional probabilities. In the implementation, we interpreted *incorrectly localized responses* as false negative or false positive. The detailed rules for matching lesions are discussed in the early sections in this paper.

Localization ROC (LROC) is an extension of ROC designed to handle locations [Starr et al., 1975]. However, LROC is not widely used, because of its limitations:

- Validation of the underlying LROC models [Wagner et al., 2002] and assumptions are still under investigation [Metz, 1996; 1999c];

- Score-keeping methods depend on the diameter of the region and method of overlap accepted for a true-positive location both in the image and in the associated pathology report [Nishikawa et al., 1998; Giger, 1996].

Therefore, LROC is not applied into our testing tool.

Our testing tool uses a continuous confidence grading scale for ROC analysis, instead of BI-RADS scale, because BI-RADS is a reporting scale designed for reporting results to patients case by case, not designed for evaluating the performance of radiologists who interpret mammograms lesion by lesion. While the BI-RADS 5 categories scale creates a significant problem for curve-fitting of "degenerate" data, our testing tool solved this problem by using a continuous scale bar and implemented an advanced ROC curve-fitting algorithm.

## V. LIMITATIONS AND FUTURE RESEARCH

The evaluation results of the testing tool also depend on the quality of test items and test construction. In the preliminary evaluation, the scores of all test takers on Test II were less than their scores on Test I, even though both tests used similar content specifications, which include the same percentage of pathologic diagnosis and finding (lesion) types. We identified that the image quality of the items in the second test was the main factor causing the difference.

In the future, a systematic evaluation of the testing tool with a large sample size of test sets and test takers should be conducted over a longer time period. A time-sequence design can be used to trace and audit residents' training in breast imaging in hospitals and medical schools. New ROC curve-fitting software will be integrated into the system. We need to explore the possibility of extending this testing tool to evaluating radiologists' performance of interpreting other medical images other than mammograms. Furthermore, future research could evaluate the ability of test

performance on the Radiology/Mammography Board Examination. More ambitiously, and ideally, the performance of physicians in their post-graduate practice could be evaluated.

**ACKNOWLEDGEMENTS**

This study was a part of the National Digital Mammogram Archive project, funded by National Library of Medicine.

*Editor's Note:* This article was received on May 2, 2005 and was published on August 8, 2005. It is one of a series of articles on decision making in health care being published by the Information Systems and Health Care department of CAIS.

**REFERENCES**

Abe, H. et al. (2003) "Computer-Aided Diagnosis in Chest Radiology: Results of Large-Scale Observer Tests Performed at the 1996-2001 RSNA Scientific Assemblies", *RadioGraphics* ,(23), pp. 255-265.

American College of Radiology (1998) *Breast Imaging Reporting And Data System*, 3rd Edition, Reston, Va: American College of Radiology.

Armato, S.G., M.B. Altman and P.J. La Rivière (2003) "Automated Detection of Lung Nodules in CT Scans: Effect of Image Reconstruction Algorithm", *Med Phys*, (30), pp. 461–472.

Bird, R.E. (1989) "Low-Cost Screen Mammography: Report on Finances and Review of 21,716 Cases", *Radiology*, (171), pp.87-90.

Esserman, L. et al. (2002) "Improving the Accuracy of Mammography: Volume and Outcome Relationships", *J. of National Cancer Institute*, 94(5), pp.369-375.

Food and Drug Administration (FDA), U.S. HHS. Federal Register (1997), *Quality mammography standards, final rule*, 21 CFR, Parts 16 and 900. P. 55852.

Giger, M.L. (1996) *Digital mammography's 96*, Amsterdam, The Netherlands: Elsevier Science, 53-59, 1996.

Hall, F.M. et al. (1988) "Nonpalpable Breast Lesions: Recommendations for Biopsy on Suspicion of Carcinoma at Mammography", *Radiology*, (167), pp.353-8.

Huber, S. et al. (1998) "Effects of a Microbubble Contrast Agent on Breast Tumors: Computer-Assisted Quantitative Assessment with Color Doppler US—Early Experience", *Radiology*, (208), pp.485-489.

Jiang, Y. et al. (1999) "Improving Breast Cancer Diagnosis with Computer-Aided Diagnosis", *Acad Radiol*, (6), pp.22-33.

Kacl, G. M. et al . (1998) "Detection of Breast Cancer with Conventional Mammography and Contrast-Enhanced MR Imaging", *Eur Radiol*, (8), pp. 194-200.

Kerlikowske, K. et al. (1993) "Positive Predictive Value of Screening Mammography by Age and Family History of Breast Cancer", *JAMA*, (270), pp.2444-2450.

Kopans, D. B. (1994) "The Accuracy of Mammographic Interpretation", *New England Journal of Medicine*, (331), pp.1521-2.

Lewin, J.M. et al. (2001) "Comparison of Full-Field Digital Mammography with Screen-Film Mammography for Cancer Detection: Results of 4,945 Paired Examinations", *Radiology*, (218)3, pp. 873-880.

Lin, J.S. et al.(1995) " Differentiation Between Nodules and End-On Vessels Using a Convolution Neural Network Architecture", *J Digit Imaging*, (8), pp. 132-41.

Linver, M.N. et al. (1992) " Improvement in Mammography Interpretation Skills in a Community Radiology Practice After Dedicated Teaching Courses: 2-year Medical Audit of 38,633 Cases", *Radiology*, (184), pp. 39-43.

Linver, M.N. (1995) " The Mammography Audit: Primer For MQSA", *AJR*, (165), pp. 19-25.

Lusted, L.B. (1971) "Signal Detectability and Medical Decision-Making", *Science,* (171), pp. 1217-1219.

MacMahon, H. Et al. (1999) "Computer Aided Diagnosis of Pulmonary Nodules: Results of A Large Scale Observer Test", *Radiology*, (213), pp.723-726.

Metz, C.E. (1986) " ROC Methodology in Radiologic Imaging", *Investigative Radiology*, (21), pp 720-731.

Metz, C.E. (1996) *Digital Mammography's 96*, Amsterdam: Elsevier Science, pp.61-68.

Metz, C.E. (1999a) "Evaluation of CAD Methods, in Computer-Aided Diagnosis in Medical Imaging", *Amsterdam: Elsevier Science*, (1182), pp. 543-554.

Metz, C.E. and X. Pan (1999b) "Proper Binormal ROC Curves: Theory and Maximum-Likelihood Estimation", *J. Math. Psych*, (43), pp.1-33.

Metz, C.E. (1999c) *Computer-aided diagnosis in medical imaging*, Amsterdam: Elsevier Science, pp.43-554.

Nishikawa, R.M. et al. (1994) "Effect of Case Selection On the Performance of Computer-Aided Detection Schemas", *Med Phys*, (21), pp.265-9.

Nishikawa, R.M. and L.M. Yarusso (1998) "Variations In Measured Performance of CAD Schemes Due to Database Composition and Scoring Protocol", *Proc SPIE*, (3338), pp. 840-844.

Poon, P.Y. et al. (1992) "Medical Audit of Mammography: a Simplified Alternative", *Canadian Association of Radiloogists Journal*, (43)3, pp.191-4.

Rockette, H.E., D. Gur and C.E. Metz (1992) " The Use of Continuous and Discrete Confidence Judgments in Receiver Operating Characteristic Studies of Diagnostic Imaging Techniques", *Investigative Radiology*, (27), pp. 169-172.

Sahiner, B. et al. (1998) "Computerized Characterization of Masses on Mammograms: the Rubber Band Straightening Transform and Texture Analysis", *Med Phys*, (25), pp.516-26.

Shusuke, Q., S. Li and K. Doi (2003) " Selective Enhancement Filters for Nodules, Vessels, and Airway Walls in Two- and Three-Dimensional CT Scans", *Med Phys*, (30),pp. 2040-2051.

Sickles, E.A. (1992) "Quality Assurance: How To Audit Your Own Mammography Pratice", *Radiol Clin North Am*, (30), pp.265-27

Spring, D.B. and K. Kimbrell-Wilmot (1987) "Evaluation the Success of Mammography at the Local Level: How to Conduct an Audit of Your Practice", *Radiol Clin North Am*, (25), 983-992.

Starr, S.J. et al. (1975) "Visual Detection and Localization of Radiographic Images", *Radiology*, (116), pp. 533-538.

Swets, J.A. (1979) " ROC Analysis Applied To the Evaluation of Medical Imaging Techniques", *Invest Radiol*, (14), pp. 109.

Wagner, R.F., S.V. Beiden and C.E. Metz (2001) " Continuous vs. Categorical Data For ROC Analysis: Some Quantitative Considerations", *Academic Radiol*, (8), pp. 328-334.

Wagner, R.F. et al (2002), "Assessment of Medical Imaging and Computer-Assist Systems: Lessons from Recent Experience", *Academic Radiol*. (9), pp. 1264-1277.

Taplin, S.H. et al. (2000) "Accuracy of Screening Mammography Using Single Versus Independent Double Interpretation", *AJR*, (174)5, pp.1257-62.

Wu**,** M., Y. Zheng and E.D. Pisano (2002) " NLM Tele-Educational System For Radiology Residents Interpreting Mammography", *Proceedings of the AMIA*, pp. 909-13.

Wu, M. and E. D. Pisano (2004) "Mammography Test Construct Tool", *AMIA 2004 Annual Symposium.*.

Zheng, Y. et al. (2004) "Online Annotation Tool For Digital Mammography"*, Academic Radiology*, 11(5), pp. 566-72.

Zweig, M.H. and G. Campbell (1993) "Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine", *Clinical Chemistry*, (39), pp.561-577.

## ABOUT THE AUTHORS

**Etta Pisano** is Professor of Radiology and Biomedical Engineering in University of North Carolina at Chapel Hill (UNC) and she is Chief of Breast Imaging at UNC Hospital.

**Min Wu** is an Assistant Professor in the healthcare informatics program of Health Sciences Department at University of Wisconsin – Milwaukee. He earned his doctorate in biomedical engineering from University of North Carolina at Chapel Hill in 2003. His research interests include medical decision making, dental informatics and biomedical database system design.

**Yuanshui Zheng** is a Research Assistant Professor in Department of Radiology in UNC. He earned his doctorate in nuclear engineering from North Carolina State University in 2002.

Communications of the Association for Information Systems

ISSN: 1529-3181