Anniversary Paper: Evaluation of medical imaging systems

Elizabeth A. Krupinski[a)]
Department of Radiology, University of Arizona, Tucson, Arizona 85724

Yulei Jiang
Department of Radiology, University of Chicago, Chicago, Illinois 60637

(Received 12 September 2007; revised 14 November 2007;
accepted for publication 10 December 2007; published 28 January 2008)

Medical imaging used to be primarily within the domain of radiology, but with the advent of virtual pathology slides and telemedicine, imaging technology is expanding in the healthcare enterprise. As new imaging technologies are developed, they must be evaluated to assess the impact and benefit on patient care. The authors review the hierarchical model of the efficacy of diagnostic imaging systems by Fryback and Thornbury [Med. Decis. Making **11**, 88–94 (1991)] as a guiding principle for system evaluation. Evaluation of medical imaging systems encompasses everything from the hardware and software used to acquire, store, and transmit images to the presentation of images to the interpreting clinician. Evaluation of medical imaging systems can take many forms, from the purely technical (e.g., patient dose measurement) to the increasingly complex (e.g., determining whether a new imaging method saves lives and benefits society). Evaluation methodologies cover a broad range, from receiver operating characteristic (ROC) techniques that measure diagnostic accuracy to timing studies that measure image-interpretation workflow efficiency. The authors review briefly the history of the development of evaluation methodologies and review ROC methodology as well as other types of evaluation methods. They discuss unique challenges in system evaluation that face the imaging community today and opportunities for future advances. © *2008 American Association of Physicists in Medicine*. [DOI: 10.1118/1.2830376]

Key words: system evaluation, medical imaging, receiver operating characteristic (ROC) analysis, diagnostic accuracy, observer study, workflow efficiency

I. INTRODUCTION

The field of medical imaging has grown immensely since Roentgen discovered x rays and realized that they could be used to look inside the human body to detect and characterize disease. Since then, diagnostic x-ray technology has evolved from film-based to completely digital where images are manipulated and viewed in a softcopy format. Advanced imaging modalities such as computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography were developed late in the 20th century and in the 21st century we witness the growth of molecular and optical imaging technologies. Radiology is not the only image-based medical specialty. Significant growth in imaging technology has been seen in pathology with the development of virtual slide processors[1] and in telemedicine with digital image acquisition for dermatology, ophthalmology, and cardiology.[2] In each of these instances, the emergence of new technologies raises important questions concerning optimization of the acquisition, storage, transfer, and display of image as well as text-based information, choice of appropriate display media and format, optimization of image compression, and optimization of image processing and computer-aided detection (CADe) and diagnosis (CADx), etc. It is only through systematic and objective evaluation of the entire imaging system—from hardware to human interpretation of images—that these questions can be answered.[3–6]

Biomedical imaging has grown so much in recent years that the National Institute of Biomedical Imaging and Bioengineering (NIBIB) was formed in 2000 as the newest institute within the National Institutes of Health.[7] Key to NIBIB's mission is "supporting studies to assess the effectiveness and outcomes of new biologics, materials, processes, devices, and procedures."[7] The crucial need for assessment of the efficacy of biomedical imaging technologies is also stressed in the recent "Blueprint for Imaging in Biomedical Research" developed jointly by the Academy of Radiology Research, the American Roentgen Ray Society and the Radiological Society of North America.[8]

The question, however, is what exactly is the purpose of evaluation? Ultimately, evaluation should be driven by a clinical question or task, which may be to detect a particular disease or to characterize some disease processes, for example. The experimental protocol and the analytical tools used to evaluate imaging results will vary depending on the nature of the clinical task. In 1991, Fryback and Thornbury[9,10] proposed a hierarchy of six levels of diagnostic efficacy, which can be used as a guiding principle in the evaluation of medical imaging systems. They defined efficacy as the probability of benefit to individuals from a system or test under ideal conditions of use, where the use of a system or test refers to the context of the clinical question or task. The six levels of efficacy they proposed are technical efficacy, diagnostic accuracy, diagnostic thinking efficacy, therapeutic efficacy, patient outcome, and societal efficacy

TABLE I. Fryback and Thornbury's proposed hierarchy (Ref. 9) of six levels of diagnostic efficacy. Efficacy is defined as benefit to individuals from a system or test under ideal conditions of use.

Levels of diagnostic efficacy	Definition	Commonly measured parameters
Technical efficacy	System or test fidelity. How accurately and precisely it measures what is to be measured.	Physical parameters (e.g., dose, DQE)
Diagnostic accuracy efficacy	How well or accurately a system or test predicts presence/absence (or extent/magnitude) of a disease or health condition.	Sensitivity, specificity, positive/negative predictive value, accuracy, ROC area under the curve (Az)
Diagnostic thinking efficacy	Impact of diagnostic test results on clinician's estimate of the probability that a patient suffers from a disease or health condition.	Changes in diagnosis, prognostic assessment, etc., before and after a diagnostic test
Therapeutic efficacy	Whether or how much the system or test changes patient's course of treatment/care.	Changes in treatment regimen—type of treatment, dose etc.
Patient outcome efficacy	Degree to which patient's health/condition improves.	Survival rates, quality of life
Societal efficacy	Impact of the system/test on society as a whole.	Cost-benefit analyses, number of lives saved

(see Table I). Briefly, technical efficacy refers to the fidelity of a system or test, or how accurately and precisely it measures what is to be measured. Methods for assessment of technical efficacy typically involve measurement of physical parameters such as the detective quantum efficiency, spatial resolution, dose, etc. For example, in the development of CT technology for breast imaging, it is important to ascertain radiation dose to the breast and compare it with that of conventional projection mammography.[11]

Diagnostic accuracy refers to how well or how accurately a system or test predicts the presence or absence of a disease or a health condition, or how well it measures the extent or magnitude of that disease or condition. Evaluation of diagnostic efficacy is a major focus of this article. It typically involves statistical figures of merit such as sensitivity, specificity, positive and negative predictive values, and the receiver operating characteristic (ROC) curve.[12] Numerous investigations addressing this level of efficacy are recent topics of interest, some of which are the evaluation of MRI for breast cancer detection and diagnosis,[13] low-dose CT for lung cancer screening,[14] and CADe and CADx tools.[15–19] Diagnostic thinking efficacy refers to the impact of diagnostic test results on the clinician's estimate of the probability that the patient suffers from an abnormality, disease, or condition. This level of efficacy can be difficult to assess, but it clearly impacts the next level of efficacy—therapeutic efficacy, which refers to whether (and how much) the system or test changes the patient's course of treatment or care. Assessment of therapeutic efficacy is the primary objective of many imaging, drug, and interventional trials. For example, a recent study on MRI of the contralateral breast in women with known breast cancer found that MRI detected four more cancers than mammography, altering the course of treatment for these patients.[13]

Patient-outcome efficacy is the most important level of efficacy from the individual patient's perspective. It measures the degree to which the patient's health or condition improves. Patient-outcome measures include such figures of merit as survival rates (often expressed in years after disease detection/treatment) and quality of life, and generally require longitudinal randomized controlled clinical trials. Interventional radiology studies often show impact of radiology on patient outcome. For example, Kim *et al.* looked at the long-term outcomes of transcatheter embolotherapy in women with chronic pelvic pain caused by ovarian and pelvic varices and found significant improvements in 83% of the patients with reduced level of pelvic pain and other symptoms as well as overall clinical improvement.[20] Molecular imaging, emerging with significant potential for measuring the outcome of individualized cancer therapy, is also an area of great interest today.[21]

The highest level of efficacy is on the society as a whole and can be very difficult to measure quantitatively. Evaluation of societal efficacy typical entails cost-benefit analyses that assess the tradeoff between costs of the system or test and benefits and savings that result—both for the individuals and society as a whole. Costs associated with performing the test are relatively easy to calculate, but other costs such as cost per life saved are more difficult to ascertain and often bring forth difficult ethical and moral issues.[22–24] For example, in January 2007, Congress cut the Medicare physician fee schedule for imaging services, citing critics who charge that some health-care providers burden society by performing more tests than necessary to boost revenue without evidence that these tests improve patient care.[25] Radiologists and imaging equipment manufacturers naturally disagree with this view and the cut.[26] Clearly, this is an issue of societal efficacy.

II. HISTORICAL PERSPECTIVE

One cannot truly consider evaluation of medical imaging systems without taking into account the entire system—from image acquisition to human interpretation of the image data, to how the diagnostic information is communicated among, and acted upon by, physicians, patients, and others. The focus on the radiologist as an integral part of the imaging system[27,28] began soon after World War II. A series of studies was conducted to determine which of four radiographic and fluoroscopic techniques was better for tuberculosis screening.[29,30] Instead of finding, as hoped, that one imaging technique was clearly superior than the others, intraobserver

and interobserver variation was found to be so large that it was not possible to determine which system was the best. A surprisingly large amount of reader variation was found even when radiologists were asked to do something as straightforward as describing the physical characteristics of radiographic shadows.[31] It became clear then that two things needed to happen: systems were needed to improve radiologists' performance and reduce their interpretation variability and methods were needed to evaluate the systems and their impact on observer performance.

In the early 1950s, progress was made in fields outside of medicine that would soon impact system and observer-performance evaluation in medical imaging. Based on principles from signal-detection theory,[32,33] ROC analysis was developed by researchers from such diverse fields as engineering, psychology, and mathematics.[34–36] The initial application was evaluation of radar operators in the detection of enemy aircraft and missiles, but its use quickly spread. Lusted first introduced ROC technique into medicine in the 1960s[37–40] and efforts to formalize medical decision-making in diagnostic medicine followed.[41–44] Since then we have witnessed a significant amount of theoretical development and practical application of ROC techniques especially in the area of radiology, facilitated to a large extent by the distribution of freely available ROC-analysis computer software.[45–47]

ROC analysis is not the only method that can be used to evaluate medical imaging systems–it is the most rigorous and the most widely accepted. We will devote much of our discussion in this article to this approach. Other approaches also have been suggested and used to varying degrees. One such approach is an experiment in which pairs of images are presented side-by-side and the observer is asked to distinguish or rank-order each image.[48–52] In recent years, this approach has been used successfully to evaluate image compression techniques, in which observers' ability to distinguish images compressed to varying degrees from the original (not-compressed) image is assessed in this type of experiments. The rationale, based on the concept of just noticeable differences (JNDs), is that if the observer is not able to distinguish an compressed image reliably from its not-compressed original, then the image compression causes only "visually lossless" changes to the image and, therefore, the changes should not affect diagnostic performance.[53–58] However, potential weaknesses of this approach include that it is subjective rather than objective assessment of observer performance and that it evaluates diagnostic performance indirectly (through the assessment of image quality). Nevertheless, this approach has been proposed as a way to plan for a large-scale ROC study—to decide whether a ROC study is justified and to help decide the number of cases and number of readers needed.

III. ROC EXPERIMENTS AND ANALYSES

Wagner, Metz, and Campbell recently published a comprehensive and in-depth review on the assessment of medical imaging systems and computer aids.[3] We encourage interested readers to consult their authoritative text; here we provide only a brief overview. The most popular ROC experiment at the present time is probably the so-called multiple-reader multiple-case (MRMC) paradigm. This experiment involves multiple cases with known disease truth status and multiple readers—most commonly every reader reads every case in every imaging modality. The intuitive rationale of the MRMC paradigm is that the performance of an imaging system is reflected in a range of case difficulty and that the imaging system is only as good as the skill of the readers who interpret the images. Thus, sampling cases of various difficulty and readers of various skill level is important for evaluation of an imaging system.

There are potential biases in the design of the MRMC experiment that one should avoid or minimize and also opportunities to make the experiment more effective or powerful.[3,59] For example, the order in which readers read images is a pertinent subject of consideration. Because there are potential advantages when a reader reads images of a patient for a second time, there is potential bias in favor of the modality that is read second compared with the modality that is read first. One way to minimize this bias is for readers to read half of the cases first in one modality and the other half of the cases first in the modality being compared. The results are subsequently combined and, therefore, any potential reading-order effect will tend to cancel out, minimizing this potential bias.[59] In most situations it is both appropriate and desirable for readers to read images of each modality independently. However, in computer-aided diagnosis, because clinically the radiologist will use the computer aid at the same time that images are interpreted, it is also appropriate in the MRMC experiment for readers to read each case first without the computer aid and then, immediately after, read the case again with the computer aid, because this study design mimics the intended clinical use of computer aid.[60] While this so-called "sequential" design is generally not appropriate for comparison of imaging modalities that are not used together clinically,[3,61] it can afford the experiment greater statistical power compared with the "independent" design in the case of computer-aided diagnosis.[62]

The types of data to collect from readers in a MRMC experiment are an important study-design consideration.[63,64] ROC analysis requires ordinal data; this is usually accomplished by asking the reader to report his or her diagnostic confidence in a specified diagnostic task. Diagnostic confidence can be expressed in terms of a 4-, 5-, or 6-point ordinal scale, or in terms of a quasicontinuous ordinal scale (e.g., 1–100).[65–67] Although the BI-RADS final assessment categories[68] have six points and have been used to estimate ROC curves,[69,70] we will return later to discuss why BI-RADS assessment categories are not appropriate for fitting ROC curves. It has been shown that if readers are able to use quasicontinuous scales, then the results can benefit ROC-curve fitting.[71] Investigation on this topic continues.[67] The report of binary action-item decisions—e.g., biopsy versus follow-up,[72] recall versus routine screening, etc.—provides

additional information on the reader's diagnostic decision that is complementary to the ROC curve[3] and provides important information for cost-benefit analyses.

A multitude of statistical methods has been developed to analyze the MRMC experiment,[3] to account for contributions to variance in the ROC curve from variations in case difficulty, reader skills, and their interactions. Swets and Pickett described the principle for analyzing the MRMC experiment,[44] and Dorfman, Berbaum, and Metz developed the first practical algorithm for this analysis.[73] Their method allows meaningful comparison of modalities, simultaneously accounting for both reader-skill variation and case-difficulty variation. Alternative methods have also been developed,[74,75] and Hillis *et al.*[76] showed recently a close relationship (or equivalence) between two of these methods.[73,74] Beiden, Wagner, and Campbell developed a method that uses bootstrapping to allow not only comparison of two modalities but also quantitative estimate of the magnitude of various variance components.[77] With their method, it is now possible to quantify explicitly the contribution of reader variability in a MRMC experiment.[78,79] Gallas later developed a different method based on an approach proposed by Barrett *et al.*[80] that provides similar estimates without invoking the method of bootstrapping.[81] It is likely that new methods for analyzing the MRMC experiment will continue to be developed in the near future.[82]

ROC analysis applies to diagnostic tasks with binary truth states, e.g., normal versus abnormal, benign versus malignant, etc. The abnormal assessment in a ROC experiment does not require location specification; rather, it is a summary assessment of the entire image or case. The location-specific receiver operating characteristic or LROC analysis,[83,84] which applies to images or cases that have only zero or one abnormality of interest, requires the observer to correctly locate the lesion in addition to correctly diagnosing it. If the observer is allowed to indicate at most one abnormal finding in each image, then the LROC curve is monotonically related to the ROC curve.[83,84] If the image may contain more than one abnormality and the observer is allowed to indicate more than one abnormal finding in each image, then the free-response receiver operating characteristic or FROC analysis[85,86] is appropriate. Breast cancer detection in screening mammography, in which radiologists often identify multiple lesions in a single image, is an example task appropriate for FROC analysis, and computer detection techniques also typically require FROC analysis. If the abnormality under study involves more than two diagnostic truth states, e.g., to differentiate solid malignant mass, solid benign mass, and cyst in the breast, than ROC analysis needs to be further generalized to multiclass ROC analysis.[87,88]

One area of recent development of ROC methods is FROC analysis.[86,89–92] Earlier FROC curve-fitting techniques did not gain widespread popularity because of concern of whether multiple observer responses made in a given image can be treated as independent. The new JAFROC method does not require this assumption. It combines FROC analysis with the method of jackknifing used by Dorfman, Berbaum, and Metz in their method[73] and simulations suggest that the JAFROC method may yield greater statistical power than other methods, including the Dorfman–Berbaum–Metz method.[73] Studies that apply the JAFROC method are beginning to appear.[91,92]

IV. EVALUATION OF OTHER HUMAN FACTORS THAT AFFECT DIAGNOSTIC PERFORMANCE

As already noted, there is a significant amount of interobserver and intraobserver variability in radiologists' diagnostic performance, and the advances made in ROC analysis in recent years can quantify much of that. An important question, however, is what in the imaging system (including the human observer) causes this variability. To help investigate this issue, other approaches to system evaluation also have been explored to investigate how the radiologist fits into, and interacts with, the imaging system. Some of these address the radiologist's working environment such as image quality, display quality, ergonomics, air quality, etc., under the assumption that if the working environment is not optimized, then diagnostic accuracy could suffer. Other methods focus on the perceptual and cognitive processes in the interpretation of medical images, with the goal of understanding how the radiologist processes image data—correctly or incorrectly. If this goal is achieved, then we can hope to optimize image quality and image presentation to better match with the human eye-brain system. We can also hope to develop computer-based tools (e.g., computer-aided diagnosis) to assist the radiologist when their perceptual or cognitive abilities tend to fail (e.g., in detecting subtle or partially obscured lesions).

Efficient interaction between human observers and imaging systems is more important today than when screen-film systems dominated. Advanced technologies such as thin-section CT, virtual colonoscopy, and MRI with various pulse sequences and contrast media combinations have resulted in thousands of images per case that the radiologist must handle. In pathology where virtual digital slides are becoming prevalent, the pathologist is also viewing larger amount of digital image data than with traditional light microscopy.[1] In both fields, the transition to digital imaging systems has resulted in significant increases in the amount of time required to view a case[1,93,94] and surveys suggest that work overload contributes substantially more to clinician dissatisfaction now than in the past.[95,96]

There is evidence that the electronic reading room leads to greater fatigue and some are investigating whether that impacts diagnostic performance.[97–99] We developed a short survey to assess radiologist fatigue, in which we asked radiologists about their symptoms of visual and physical fatigue, the types (film, digital, or both), modality, and number of cases interpreted, and the total reading time.[100] The survey was given to attending and resident radiologists in the Radiology Department at the University of Arizona at various time of the day over several days. Table II and Fig. 1 present correlations between symptoms of fatigue with reading time and the number of cases read. For all symptoms except for headache and shoulder strain, there was a significant positive

TABLE II. Correlation between subjective fatigue, and reading time and the number of cases read.

Symptoms of fatigue	Correlation with reading time	Correlation with the number of cases read
Blurred vision	$R=0.344$ $p=0.0113$	$R=0.422$ $p=0.0015$
Eyestrain	$R=0.429$ $p=0.0012$	$R=0.475$ $p=0.0003$
Difficulty focusing	$R=0.384$ $p=0.0042$	$R=0.446$ $p=0.0007$
Headache	$R=0.235$ $p=0.0899$	$R=0.432$ $p=0.0011$
Neck strain	$R=0.384$ $p=0.0042$	$R=0.549$ $p<0.0001$
Shoulder strain	$R=0.250$ $p=0.0711$	$R=0.469$ $p=0.0003$
Back strain	$R=0.304$ $p=0.0265$	$R=0.424$ $p=0.0014$
General fatigue	$R=0.471$ $p=0.0003$	$R=0.642$ $p<0.0001$

correlation between the reported symptoms of fatigue and reading time, and for all symptoms there was significant positive correlation between the reported symptoms of fatigue and the number of cases read. This suggests that in the future development of display systems (i.e., computer workstations) for routine clinical use, attention needs to be directed to the comfort and physical wellbeing of the radiologist.

Measuring the time it takes for a radiologist to render a diagnostic decision using an imaging system is also an important evaluation tool. In today's high-volume reading environment, it is particularly important to investigate how imaging systems can be optimized to reduce interpretation time. For example, CADe tools are being integrated rapidly into a number of digital imaging modalities with the dual goals of improving diagnostic accuracy and decreasing interpretation time. Halligan *et al.* investigated the latter goal by comparing CADe versus no CADe for CT colonography and found that interpretation time decreased significantly with CADe (2.9 min versus 1.9 min).[101] Similarly, temporal subtraction (13.6 s per case without subtraction versus 10.8 s with subtraction)[102] and stack mode presentation of multi-slice image data (75 s per case for tile mode versus 63 s for stack mode)[103] significantly reduce interpretation time.

Viewing time can be measured simply with a stopwatch[104] or sophisticatedly with computer auditing tools that automatically record every interaction between the radiologist and a workstation.[105,106] Another sophisticated evaluation method is eye-position recording (see Fig. 2).[4,27,107] Eye-tracking studies have been used to gain basic understanding of the visual search and decision-making process[108–112] and also for system evaluation.[113] Although typically done in dedicated image-perception laboratories, eye-tracking studies are useful in general to understand how an imaging system affects interpretation efficiency and the decision-making process. Whereas ROC analysis assesses the final decision, eye-tracking studies provide information on how the observer

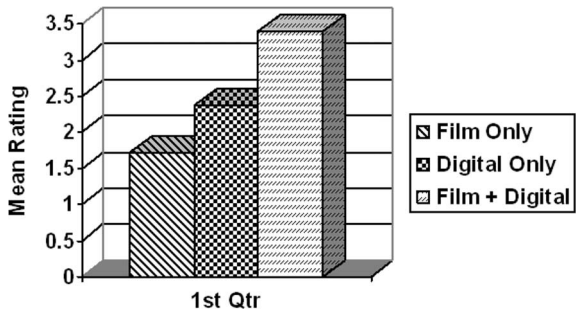
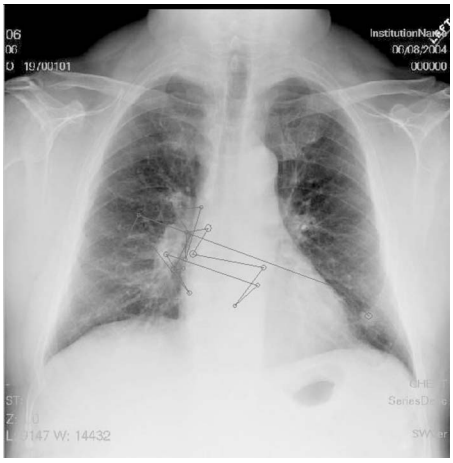
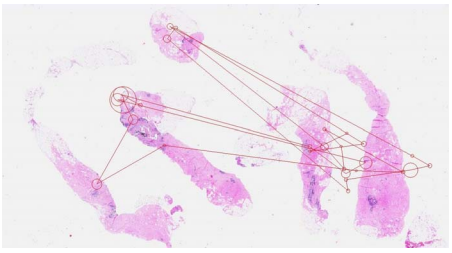


FIG. 1. Average survey ratings of the symptom of "blurred vision" as a function of the types of images (film only, digital only, and both film and digital) read in a single day.



(a)



(b)

FIG. 2. (a) Example of a typical eye-position pattern generated by an experienced radiologist. The small circles represent fixations or where the high-resolution portion of the gaze lands and the lines represent the order in which they were generated. The task was to search for lung nodules. In this case there is a nodule in the lower left lung. The radiologist stopped searching the image once a fixation landed here and he correctly reported the nodule as present. (b) Example of a typical eye-position pattern generated by an experienced pathologist. The small circles represent fixations or where the high-resolution portion of the gaze lands and the lines represent the order in which they were generated. The task was to select the three areas the observer would want to magnify to view diagnostic tissue at a higher resolution.

reaches that decision. A major assumption behind eye-tracking studies is that the amount of time spent looking at features in the image reflects information processing, object encoding, and recognition. By correlating eye-position parameters such as dwell time, number of returns to a location, and saccade length (i.e., hops between fixations) with various (true positive, false positive, true negative, and false negative) decisions, it is possible to draw conclusions about perceptual and cognitive processes that are the foundation of image interpretation.

For example, Krupinski *et al.* carried out a series of studies on the influence of various properties of digital image display on visual search and decision-making efficiency. Three basic parameters characterize visual search efficiency: total viewing time, time to first fixation on a lesion of interest with high-resolution foveal vision, and cumulative time spent on the lesion that can be correlated with a decision. Total viewing time is the time the observer spends looking at the image. Time to first fixation on a lesion refers to the time it takes for the observer to come to the first fixation on a lesion of interest. Cumulative time (also referred to as cumulative decision dwell time) for true positives and false negatives is calculated by defining a region of useful visual field (i.e., the extent of high resolution vision; typically 2.5° radius) centered on the lesion and summing the time associated with all fixations within this region of useful visual field. If the observer task involves explicit localization, then false-positive cumulative decision dwell time can be calculated also in this fashion with the region of useful visual field centered on the image location identified by the observer. True-negative decision dwell time is calculated from image regions that are lesion free but receive fixation clusters.[114]

Higher display luminance,[115] calibration with the Digital Imaging and Communications in Medicine (DICOM) Grayscale Display Function standard,[116] high-performance monochrome rather than color display,[117] and 11-bit rather than 8-bit display[118] all result in higher search efficiency and improved diagnostic accuracy. Even the graphical user interface (GUI) and the way in which images and tool bars are positioned on the display can affect search efficiency. A study of softcopy reading of bone images showed that radiologists spent 20% of their interpretation time looking at nondiagnostic menus and tool bars.[119] Understanding how the GUI affects users accessing information is useful for both GUI designers and radiologists.[100]

ROC, workflow, and eye-tracking studies require large amounts of time and resources whether it is a small-scale laboratory study or a large-scale clinical trial. Getting enough images with known "gold-standard" diagnostic truths and enough observers to attain sufficient statistical power can be daunting. This is particularly true when multiple conditions need to be studied because the conditions that most likely impact observer performance are difficult to predict *a priori*. For example, as images become larger and numbers of images increase, image compression may become a necessity rather than option.[120] But one size does not fit all—e.g., the image-compression ratio that works for CT abdomen may not work for CT chest. As already noted, an alternative to the ROC study that is gaining popularity for system evaluation tasks such as determining the appropriate level of image compression is the concept of the visually lossless threshold.[53–58]

A useful approach to system evaluation that does not require human observers is modeling.[6,121,122] There are a number of models of human vision that predict human detection performance and those based on the concept of JNDs appear to simulate human performance closely.[122,123] The idea is to input a pair of images (e.g., one displayed on a LCD and one on a CRT) into the model which yields a JND map of the magnitude and spatial location of visible differences between the images. Various stages of the model simulate everything from the optics of the eye to a phase-independent energy response that mimics the transformation that occurs in the mammalian visual cortex from a linear response of simple cells to an energy response of complex cells.[107] These models often correlate well with human performance and have been used to evaluate imaging system components such as phosphor for CRT display,[124] image window/level over an entire image versus over a small region covering a lesion,[125] the effect of display veiling glare on performance,[126] and reconstruction algorithms for parallel MRI with multiple coils and *k*-space subsampling.[122] However, it is often necessary to carry out at least limited human observer studies with qualified observers and appropriate images to validate model predictions. While promising and interesting, to date this approach has had only limited success in but a few simple clinically applicable scenarios.

V. CHALLENGES AND OPPORTUNITIES

In performing ROC analysis today, researchers face the challenge of rapid development of new methodologies that are not always accompanied by publicly available, reliable, and easy to use computer software. Three groups—the University of Chicago, the University of Iowa, and the University of Pittsburgh—consistently post updated software.[45–47] Many other computer programs are either out-of-date, difficult to use, or not reliable. For researchers to take advantage of advanced ROC methodologies, software tools need to be updated, tested, and made available on a regular basis. A particular need is computer software for statistical power analysis of ROC studies. Only the University of Chicago and the University of Iowa provide software for power calculation and sample-size analysis, with examples from the literature.[45,46] There are other useful guidelines and tables in the literature, which do not have accompanying computer programs.[127–130]

Another challenge is between laboratory evaluation studies, clinical trials, and translation to clinical use. Laboratory studies commonly use a larger proportion of abnormal cases (typically about half of all cases[131,132]) than in clinical practice to maximize statistical power; therefore, disease prevalence often does not accurately reflect clinical reality. Disease prevalence alters observers' expectations and could affect their threshold for calling a suspicious feature a lesion. Evidence suggests that disease prevalence can also affect ob-

servers' confidence ratings. Gur *et al.* studied five prevalence levels of abnormalities in chest images (nodules, interstitial disease, and pneumothorax) and found that confidence ratings tend to be higher with low disease prevalence.[133] Does the change in disease prevalence affect radiologists' diagnosis performance? Are radiologists able to detect more or fewer breast cancers at altered cancer prevalence? The studies of Gur *et al.* indicate that observer performance is not affected by change in disease prevalence.[133] However, after accounting for differences in disease prevalence do radiologists also operate at the same operating point as they do in clinical practice? Do they overcall or undercall in laboratory tests? How is performance affected by the conscious knowledge of laboratory tests not affecting patient care and the apparent similarity of laboratory tests to competitive-test environment rather than to clinical practice? ROC experiments generally study one abnormality at a time, but multiple diagnostic findings are common in clinical practice. Limiting the study to a single abnormality does not represent clinical reality accurately, and more importantly ignores the phenomenon of satisfaction of search, in which the detection of one abnormality precludes the detection of additional abnormalities.[134–136] Most laboratory studies also do not include clinical history, previous image, and data such as clinical laboratory report. Studies show that clinical history can improve diagnostic accuracy,[137] therefore withholding clinical history likely causes underestimation of diagnostic performance.

Opportunities abound. New technologies, such as breast CT, virtual colonography, molecular imaging, optical imaging, and radionuclide imaging, that push the boundary of our understanding of the biological processes underlying human health and disease are being explored.[8] Image reconstruction and analysis, computer-aided detection and diagnosis, multimodality comparison and integration, and a host of other software tools are needed to help clinicians make sense of the image data and render the best diagnostic decision. Characterization of the impact of these new technologies and tools on the daily clinical routine, e.g., human-computer interface, ergonomics, and impact on decision-making, is a fast growing area of evaluation. Clinical radiology and other specialties have only begun adopting these methods and technologies—not as pieces of hardware or software, but as integrated systems that include the human observer in a complex environment. As the digital reading environment becomes more complex, and as physical and psychological problems such as carpal-tunnel syndrome[138] and visual and physical fatigue begin to emerge,[100,139] we need to evaluate imaging systems not only with respect to diagnostic accuracy, but also toward the totality of perceptual, cognitive, and environmental factors that contribute to the diagnostic decision-making process.

VI. HOW WELL DO LABORATORY STUDIES PREDICT CLINICAL PERFORMANCE?

We began this review by citing the six tiers of diagnostic efficacy of Fryback and Thornbury as a guiding principle for system evaluation.[9,10] Implicit in this principle is the possibility that efficacy at a lower-tier level does not necessarily imply efficacy at a higher level. This is an unfortunate possibility that researchers must confront as new imaging systems are developed. In the following we discuss some aspects of the relationship between laboratory studies and clinical performance.

VI.A. Laboratory tests, field tests, and mortality trials

To put system evaluation into the framework of Fryback and Thornbury's six tiers of diagnostic efficacy,[9] it is necessary to distinguish laboratory tests, field tests, and mortality trials. We call the typical laboratory observer study[3,59] as "lab test" because these experiments involve cases with known diagnostic "truth" and experiment in the laboratory with clinical radiologists. The objective of the laboratory test is to measure or compare clinical diagnostic capability of imaging technologies for a specified diagnostic task, but it is done in the laboratory with retrospective reading of cases by observers who are aware that their image interpretation does not impact patient care. Clearly there are differences between laboratory test and clinical use of an imaging technology—we have already raised some questions that concern whether specific lab-test results accurately correlate with benchmark performance in clinical practice.

Field tests are evaluations of imaging technology in the clinical setting. Field tests are often done when a new imaging technology is first introduced into clinical practice or, at a later time, to reassess the efficacy of an imaging technology. Cancer detection rate and some form of the false-positive rate are common end points in cancer-related imaging trials.[140–142] Common end points of breast cancer screening trials are cancer detection rate—the number of cancers diagnosed per 1000 women screened—and recall rate—the proportion of women in screening recalled for diagnostic imaging study. These end points are readily measurable. Although sensitivity (the fraction of patients with cancer correctly diagnosed) and specificity (the fraction of patients without cancer correctly diagnosed) are more informative, they require complete ascertainment of whether cancer is present in every patient, which is an extremely difficult task. The randomized controlled trial is a cornerstone of medical field tests, particularly for drug and interventional procedures. However, because imaging a patient with one modality usually does not preclude imaging the patient again with another modality, imaging trials can be designed differently from, and more efficiently than, the standard randomized controlled trials. There are two common designs for imaging trials. In the first, each patient is imaged with two imaging modalities and comparison is made in the same patient cohort. Each patient serves as his or her own control. For example, in the Digital Mammographic Imaging Screening Trial both screen-film and full-field digital mammograms were obtained in each patient and the diagnostic performance of radiologists reading the two mammograms was compared.[143,144] In Freer and Ulissey's study of CADe, they first read each case without the computer aid and then, after

recording the film-only finding, read the case again with the computer aid. They then compared the film-only findings with the CADe-findings.[141] We call this type of imaging trial the "head-to-head" comparison.

In the second type of imaging trial, two imaging technologies are compared in different patient cohorts—typically the performance of a new imaging technology in a current patient cohort is compared with the performance of a standard imaging technology in a previous cohort study. The previous cohort serves as control of the current cohort. For example, to compare CADe with reading mammograms without computer aid, Gur *et al.* compared screening mammography performance before (January 1, 2000–June 30, 2001) and after (October 1, 2001–December 31, 2002) the installation of a CADe device at their institution.[142] We call this type of imaging trial the "historical-control" study. Both types of trials have advantages, disadvantages, and potential biases. In a head-to-head CADe trial, the performance measurement of the first read can be potentially biased because the radiologist could be either less vigilant than usual and rely on the additional second read to catch more cancer, or more vigilant than usual if the radiologist try to "beat" the computer. On the other hand, in a historical-control trial, one cannot distinguish the effects of differences in imaging technologies from the effects of longitudinal changes in disease prevalence, radiologists' performance, etc. A head-to-head comparison is statistically more powerful than a historical-control study because in a head-to-head comparison statistical variations tend to be matched, to some degree, in the two modalities, making it easier to observe their differences; whereas in a historical-control study statistical variations are independent in the two arms, making it difficult to observe difference between two modalities.

Mortality trials compare the number of deaths from a particular disease in a patient cohort that participates in an imaging study (trial group) with the number of deaths in another cohort that does not participate in the imaging study (control group). The objective is to answer the question: does the imaging technology save lives?[145–152] Mortality trials are highly important and perhaps the most important for the individual patient because cancer detection does not always cause a reduction in cancer mortality. For example, detection of late-stage, advanced cancer may not reduce cancer mortality, but detection of small, early-stage cancer often does. However, mortality trials by necessity are almost always randomized controlled trials that require extremely large number of patients, decades of follow-up, huge demand on resources, and raise potentially difficult ethical questions of assigning individuals to the control group when the prevailing assumption is that screening benefits them. For these reasons, field tests—not mortality trials—are often more appropriate to evaluate new imaging technologies.

VI.B. Higher ROC curves and increased cancer detection rate: Some idealized considerations

Does the lab-test result of higher ROC curves necessarily predict greater cancer detection rate in field tests? To answer this question, let us consider three highly idealized scenarios. First, let us consider a hypothetical new imaging technology that is as capable as the standard-of-care imaging technology at detecting cancer of every type. In this situation, laboratory tests will likely find the two technologies share similar ROC curves and field tests will likely find the technologies have similar cancer detection rates (though because of statistical sampling variations, some studies may find the new technology with higher cancer detection rate, others vice versa, and still others fail to find differences—the overall conclusion is, therefore, that the technologies are similar).

Let us consider another hypothetical new imaging technology that is able to detect cancer of every type that the standard-of-care imaging technology can detect, but the new technology detects the cancer earlier—when the cancer is smaller and less conspicuous. In this situation, if cases that show an advantage of the new technology are studied, laboratory tests will likely find the new technology to be associated with higher ROC curves when it is compared with the current technology. Will the cancer detection rate increase in field tests? Because the new technology detects cancer earlier, more cancers will be detected when the new technology is first put into clinical service. However, as time goes on, the increase in cancer detection rate cannot be sustained because after the new technology detects more cancers early on, fewer cancers will be there waiting to be detected in subsequent screening rounds (barring unrelated opportune increase in the underlying incidence of cancer). Over time, a steady state will commence in which the cancer detection rate of the new technology will approximately equal that of the current technology in comparable patient cohorts, but the new technology detects more small and early cancers than the current technology.[153] Therefore, an initial transient period of increased cancer detection rate may appear when the new technology is introduced clinically—only to disappear later. Even the possibility of a transient increase in the cancer detection rate will be uncertain because it will be affected by many factors such as the new technology being adopted at different times and at a different pace by different clinical groups, the learning curve of the new technology may vary for individual radiologists, and patient demographics (such as willingness to participate in screening[154]) may change over time. However, regardless of whether an increase in cancer detection rate occurs and even though in the long term sustained increase in cancer detection rate is not expected, a new imaging technology that detects cancer earlier should lead to a reduction in cancer mortality if cancer size at detection correlates with cancer mortality.[155]

Let us consider a third hypothetical new imaging technology that detects new types of cancer that the standard-of-care imaging technology is unable to detect. In this situation, if the new cancer types are studied, laboratory tests likely will find the new technology to be associated with higher ROC curves when it is compared with the current technology. Sustained increase in cancer detection rate may also occur if the new cancer types count toward the cancer detection rate. However, whether the increased cancer detection is justified or desirable will depend on whether detection of the new

types of cancer reduces cancer mortality. If an imaging technology detected interval breast cancers—fast-growing cancer that becomes clinically evident between successive mammogram screening rounds—then a mortality benefit would be likely if the cancer were detected early enough to be arrested before it causes death. However, if a new imaging technology detected indolent cancers—slow-growing cancer that patients die with rather than die from—then mortality reduction would not be likely. We have seen persistent controversies in screening mammography[156–159] (and lung[147,148] and prostate cancer screening) regarding whether the rise in the number of cancers detected from screening corresponds to the detection of cancers that kill or cancers that are indolent. A related current debate concerns the increased detection of *in situ* cancers of the breast from screening mammography and possibly enhanced with CADe.[70,160] The natural history of each type of cancer will ultimately decide whether the detection of new cancer types should count toward cancer detection rate and whether the resulting increase in cancer detection rate is justified. (These issues are also known as overdiagnosis, such as the detection of indolent cancers, and lead-time bias, which refers to credit inappropriately attributed to a screening method that detects cancer early but does not reduce cancer mortality because detection of the particular cancers early has no effect on the cancers' impact on mortality.)

These highly idealized considerations suggest that whether higher ROC curves in laboratory tests indicate higher cancer detection rate in field tests and/or reduced mortality depends on the types and natural history of cancer detected with a candidate imaging technology—and many other important factors. It is possible for a single new imaging technology to embody all three of these scenarios—detecting some types of cancer earlier than the current technology and detecting some new types of cancer that the current technology is not able to detect, but detecting other types of cancer as well as—or not as well as—the current technology. In this more complex situation, whether there is a connection between higher ROC curves in laboratory tests and higher cancer detection rate in field tests will depend on the relative weighting of the individual effects of different types of cancer—and will be associated with greater uncertainty.

VI.C. Some practical considerations

In our discussion so far of cancer detection rates, we have carefully avoided discussing "observing" an increase in the cancer detection rate in field tests because observing an actual increase in the cancer detection rate adds yet another layer of complexity to this already complex subject. Cancer is a rare event in many screening situations.[161] For every 1000 asymptomatic and average-risk women screened for breast cancer in the United States only about five breast cancers are detected by any method.[162,163] The measurement of a 0.5% cancer detection rate is associated, unavoidably, with large statistical uncertainty. Another important contributor of statistical uncertainty is inter-radiologist variability:[164,165] each individual radiologist may operate at different cancer-detection rates. Although measuring the combined cancer detection rate of a group of radiologists is statistically more reliable, statistical and inter-radiologist variability will still affect the cancer detection rate. Based on data from over two million screening mammograms read by 510 radiologists in seven U.S. regions from 1996 to 2002 (part of the Breast Cancer Surveillance Consortium[163]), Jiang *et al.* estimated that if a hypothetical new technology consistently allows each radiologist to detect one additional cancer per 1000 screening examinations compared with screening mammography (which operates at 77% sensitivity or detecting 3.94 cancers per 1000 screening examinations[162])—a very large, 25%, increase in the cancer detection rate—then the minimum required size of a field test to attain 80% statistical power to detect higher cancer detection rate is 25 radiologists each reading at least 8000 cases (200 000 patients), or 91 radiologists each reading 1000–2000 cases (91 000–182 000 patients). These are very large trials, and a larger sample of radiologists can afford a trial a smaller patient cohort—indicating the strong effect of inter-radiologist variability. Smaller trials suffer from the risk that one could observe lower cancer detection rate than the standard of care—completely opposite to the large postulated increase in the cancer detection rate.[162]

This discussion focuses on statistical power and the effect of inter-reader variability. There are many other practical issues impacting the results of trials that we have not discussed. For example, a small amount of data contamination—wherein the interpretation result of one modality is incorrectly attributed to the competing modality—often cannot be avoided completely. In another example, there are numerous sources of potential biases in statistical analysis, one of which is potential bias in the diagnostic truths. For example, the ascertainment of diagnostic truths often cannot be separated completely from the current imaging technology. If a new imaging technology makes it possible to detect smaller cancers earlier than the current technology, then a historical-control comparison of the sensitivity of this new technology with that of the current technology can be biased because the small cancers are counted in the sensitivity of the new technology but not counted in the sensitivity of the current technology as they are not detected, and therefore not known, in the current-technology arm. These and many other factors make it difficult to ascertain the true effects in a trial.

The lack of a clear link between higher ROC curves in laboratory tests and better cancer detection performance in field tests[166,167] presents substantial challenges—to the medical imaging community, the broader medical community, public policy stakeholders, the insurance industry, and the general public. Any reasonable person would expect that a better technology proven in the laboratory will also perform better in the clinic; yet there are many reasons that this expectation may not bear out. We cannot ignore the possibility of not being able to ascertain consistently the superiority of better imaging technologies in clinical settings unless we resort to extremely large trials.[29,162] Computer-aided detection of breast cancer in screening mammograms is an example.

After years of development, laboratory studies showed that radiologists operate on higher ROC curves when they are assisted by the computer compared with reading mammogram alone.[131,132,168–173] There is also a body of literature that shows the potential benefit of CADe in screening mammography[174–183] and of similar computer aids in other diagnostic tasks.[60,72,184–189] After CADe devices are introduced clinically, head-to-head comparisons and one historical-control study of CADe versus radiologists' reading mammogram alone found CADe associated with increased cancer detection and increased recall rate.[141,190–196] However, the two by far largest field tests—both historical-control studies—found little or no increase in cancer detection from CADe.[70,142,197] Although these two large studies lacked the statistical power needed to detect an increase in the cancer detection rate if it were as large as Jiang *et al.* postulated in their study,[162] conclusions were nonetheless drawn suggesting that computer-aided detection is not associated with improved detection of breast cancer.[70] These contradictory results from laboratory tests, head-to-head field tests, and larger historical-control field tests remain the subject of current interest and debate.[160,198,199] Although this debate focuses on the important question of whether computer-aided detection improves cancer detection, our inability to find a clear answer to this question may unfortunately influence the development and use of this new technology more than the truth itself.

Perhaps one reason for this disconnect between laboratory tests and field tests is that the sophisticated statistical methodologies[3] developed for ROC experiments are not used in field tests, which rely on less powerful statistical methods. Therefore, a possibility for bridging laboratory tests and field tests is to bring ROC methodology into field tests. If, during clinical image interpretation, the radiologist could provide ROC-type data as in laboratory MRMC experiment—diagnostic confidence ratings in addition to binary action-type decisions (e.g., recall versus routine screening) —then a ROC curve can be constructed subsequently when the truth status of the cases becomes available through follow-up. In this way, ROC analysis will become available to field studies. There are some early examples of this kind of study.[70,144] However, aside from a host of methodological issues that must be addressed, it is likely that one must overcome a cultural barrier in clinical radiology where binary action-type decisions are the mainstay. Clinical radiologists need to be convinced that quantitative diagnostic assessments provide richer diagnostic information that allows for the estimation of ROC curves, which in turn can provide valuable feedback to them to improve diagnostic performance. There are many methodological challenges. For example, radiologists are now accustomed to the BI-RADS final assessment categories[68] and these scales have been used as a basis for ROC analysis.[69,70] However, fundamental questions can be raised concerning the BI-RADS categories because they do not provide an ordinal scale—a fundamental assumption in ROC analysis. BI-RADS rating 2 (benign abnormality) does not imply greater suspicion of cancer than BI-RADS rating 1 (no abnormality) and BI-RADS rating 0 (incomplete study) does not imply less suspicion than any other BI-RADS ratings. BI-RADS ratings 3, 4, and 5 are not intended for screening studies.[200] For radiologists who use only ratings 0, 1, and 2, the scale reduces to three points and produces only two points in the interior of the ROC plot. Other radiologists who use the entire 6-point scale in effect use a difference rating scale, which raises questions for combining the rating data with those from radiologists who use the 3-point scale.

Currently, MRMC analysis is applied most often to experiments in which every patient is imaged in every modality and every reader reads every case in every modality. This design provides the greatest statistical power.[3] However, in principle,[44,201] MRMC analysis can be applied to multiple-reader multiple-case data in which each patient is imaged only once in a single modality and each case is read only once by a single reader—or any variant of that situation, e.g., some or all patients are imaged in more than one modality; some or all readers read cases in more than one modality; some or all cases are read several times by one reader or by several readers; etc. This more general view of the MRMC paradigm is applicable to ROC analysis in clinical practice, where it is probably not possible to obtain MRMC data as in conventional laboratory studies where every reader reads every case in every modality. However, fundamental modifications of current MRMC-analysis methods[73–75,77,81] must be made first—the feasibility of which is not entirely clear at this time—before such MRMC analysis of clinical data becomes possible.

VII. SUMMARY

Clearly, system evaluation is a multifaceted process that can be approached from a variety of perspectives. However, there has been a considerable amount of methodological development and innovation to carry out statistical analysis in the evaluation of medical imaging systems. Progress in system evaluation has paralleled progress in technological system developments and both will continue to be developed and refined. Through continuing medical imaging system development and system evaluation, diagnostic accuracy by both humans and computers will continue to improve and positively impact patient care.

ACKNOWLEDGMENTS

We are deeply grateful to Dr. Charles E. Metz, Dr. Robert M. Nishikawa, and Dr. Diana L. Miglioretti for many invaluable discussions in the preparation of this article. Y.J. is supported in part by the NCI/NIH through Grant No. R01 CA92361. E.K. is supported in part by the NIH/NIBIB through Grant Nos. R01 EB004987 and R01 EB008055.

[a]Author to whom correspondence should be addressed. Present address: Department of Radiology, P.O. Box 245067, University of Arizona, Tucson, AZ 85724. Telephone: 520-626-4498; Fax: 520-626-4376. Electronic mail: krupinski@radiology.arizona.edu

[1]R. S. Weinstein *et al.*, "An array microscope for ultrarapid virtual slide processing and telepathology. Design, fabrication, and validation study," Hum. Pathol. **35**, 1303–1314 (2004).

[2]E. Krupinski, M. Nypaver, R. Poropatich, D. Ellis, R. Safwat, and H. Sapci, "Clinical applications in telemedicine/telehealth," Telemed. J. **8**,

13–34 (2002).
[3]R. F. Wagner, C. E. Metz, and G. Campbell, "Assessment of medical imaging systems and computer aids: A tutorial review," Acad. Radiol. **14**, 723–748 (2007).
[4]D. J. Manning, A. Gale, and E. A. Krupinski, "Perception research in medical imaging," Br. J. Radiol. **78**, 683–685 (2005).
[5]E. A. Krupinski, S. Dimmick, J. Grigsby, G. Mogel, D. Puskin, S. Speedie, B. Stamm, B. Wakefield, J. Whited, P. Whitten, and P. Yellowlees, "Research recommendations for the American Telemedicine Association," Telemed. J. **12**, 579–589 (2006).
[6]H. H. Barrett and K. J. Myers, *Foundations of Image Science* (Wiley, Hoboken, NJ, 2004).
[7]National Institute of Biomedical Imaging and Bioengineering (NIBIB), http://www.nibib.nih.gov/About/MissionHistory, last checked June 7, 2007.
[8]R. L. Ehman, W. R. Hendee, M. J. Welch, N. R. Dunnick, L. B. Bresolin, R. L. Arenson, S. Baum, H. Hricak, and J. H. Thrall, "Blueprint for imaging in biomedical research," Radiology **244**, 12–27 (2007).
[9]D. G. Fryback and J. R. Thornbury, "The efficacy of diagnostic imaging," Med. Decis. Making **11**, 88–94 (1991).
[10]J. R. Thornbury, D. G. Fryback, R. A. Goepp, L. B. Lusted, K. I. Marton, B. J. McNeil, C. E. Metz, and M. C. Weinstein, *NCRP Commentary No. 13—An Introduction to Efficacy in Diagnostic Radiology and Nuclear Medicine* (National Council on Radiation Protection and Measurements, Bethesda, MD, 1995).
[11]J. M. Boone, A. L. Kwan, K. Yang, G. W. Burkett, K. K. Lindfors, and T. R. Nelson, "Computed tomography for imaging the breast," J. Mammary Gland Biol. Neoplasia **11**, 103–111 (2006).
[12]C. E. Metz, "Receiver operating characteristic analysis: A tool for the quantitative evaluation of observer performance and imaging systems," J. Am. Coll. Radiol. **3**, 413–422 (2006).
[13]C. D. Lehman, J. D. Blume, D. Thickman, D. A. Bluemke, E. Pisano, C. Kuhl, T. B. Julian, N. Hylton, P. Weatherall, M. O'loughlin, S. J. Schnitt, C. Gatsonis, and M. D. Schnall, "Added cancer yield of MRI in screening the contralateral breast of women recently diagnosed with breast cancer: Results from the International Breast Magnetic Resonance Consortium (IBMC) trial," J. Surg. Oncol. **92**, 9–15 (2005).
[14]New York Early Lung Cancer Action Project Investigators, "CT screening for lung cancer: Diagnoses resulting from the New York Early Lung Cancer Action Project," Radiology **243**, 239–249 (2007).
[15]M. Freedman and T. Osicka, "Reader variability: What can we learn from computer-aided detection experiments," J. Am. Coll. Radiol. **3**, 446–455 (2006).
[16]K. Doi, "Current status and future potential of computer-aided diagnosis in medical imaging," Br. J. Radiol. **78**, S3–S19 (2005).
[17]K. Awai, K. Murao, A. Ozawa, Y. Nakayama, T. Nakaura, D. Liu, K. Kawanaka, Y. Funama, S. Mirishita, and Y. Yamashita, "Pulmonary nodules: Estimation of malignancy at thin-section helical CT—Effect of computer-aided diagnosis on performance of radiologists," Radiology **239**, 276–278 (2006).
[18]Q. Li, F. Li, K. Suzuki, J. Shiraishi, H. Abe, R. Engelmann, Y. Nie, H. MacMahon, and K. Doi, "Computer-aided diagnosis in thoracic CT," Semin. Ultrasound CT MR **26**, 357–363 (2005).
[19]K. Horsch, M. L. Giger, C. J. Vyborny, L. Lan, E. B. Mendelson, and R. E. Hendrick, "Classification of breast lesions with multimodality computer-aided diagnosis: Observer study results on an independent clinical data set," Radiology **240**, 357–368 (2006).
[20]H. S. Kim, A. D. Malhotra, P. C. Rowe, J. M. Lee, and A. C. Venbrux, "Embolotherapy for pelvic congestion syndrome: Long-term results," J. Vasc. Interv. Radiol. **17**, 289–297 (2006).
[21]D. A. Mankoff, F. O'Sullivan, W. E. Barlow, and K. A. Krohn, "Molecular imaging research in the outcomes era: Measuring outcomes for individualized cancer therapy," Acad. Radiol. **14**, 398–405 (2007).
[22]W. Hollingworth and D. E. Spackman, "Emerging methods in economic modeling of imaging costs and outcomes: A short report on discrete event simulation," Acad. Radiol. **14**, 406–410 (2007).
[23]A. Z. Kielar, R. H. El-Maraghi, and R. C. Carlos, "Health-related quality of life and cost-effectiveness analysis in radiology," Acad. Radiol. **14**, 411–419 (2007).
[24]B. J. Hillman, "Health services research of medical imaging: My impressions," Acad. Radiol. **14**, 381–384 (2007).
[25]U.S. Department of Health and Human Services Centers for Medicare and Medicaid Services, http://www.cms.hhs.gov/PhysicianFeeSched/, last accessed June 15, 2007.
[26]M. Perrone, "MRI, x-ray firms fight Medicare cuts," Associated Press, June 6, 2007.
[27]H. L. Kundel, "History of research in medical image perception," J. Am. Coll. Radiol. **3**, 402–408 (2006).
[28]E. A. Krupinski, "The future of image perception in radiology: Synergy between humans and computers," Acad. Radiol. **10**, 1–3 (2003).
[29]C. C. Birkelo, W. E. Chamberlain, and P. S. Phelps, "Tuberculosis case finding. A comparison of the effectiveness of various roentgenographic and photofluorographic methods," JAMA, J. Am. Med. Assoc. **133**, 359–366 (1947).
[30]L. H. Garland, "On the scientific evaluation of diagnostic procedures," Radiology **52**, 309–328 (1949).
[31]R. R. Newell, W. E. Chamberlain, and L. Rigler, "Descriptive classification of pulmonary shadows. Revelation of unreliability in roentgenographic diagnosis of tuberculosis," Am. Rev. Tuberc. **69**, 566–584 (1954).
[32]A. Wald, *Statistical Decision Functions* (Wiley, Inc., New York, 1950).
[33]W. W. Peterson, T. L. Birdsall, and W. C. Fox, "The theory of signal detectability," IEEE Trans. Inf. Theory **4**, 171–212 (1954).
[34]W. P. Tanner and J. A. Swets, "A decision-making theory of visual detection," Psychol. Rev. **61**, 401–409 (1954).
[35]D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics* (Krieger, Huntington, NY, 1974).
[36]J. P. Egan, *Signal Detection Theory and ROC Analysis* (Academic, New York, 1975).
[37]L. B. Lusted, "Logical analysis in roentgen diagnosis," Radiology **74**, 178–193 (1960).
[38]L. B. Lusted, *Introduction to Medical Decision Making* (Charles C. Thomas, Springfield, IL, 1968).
[39]L. B. Lusted, "Perception of the Roentgen image: Applications of signal detection theory," Radiol. Clin. North Am. **7**, 435–459 (1969).
[40]L. B. Lusted, "Signal detectability and medical decision making," Science **171**, 1217–1219 (1971).
[41]B. J. McNeil and S. J. Adelstein, "Determining the value of diagnostic and screening tests," J. Nucl. Med. **17**, 439–448 (1976).
[42]B. J. McNeil and J. A. Hanley, "Statistical approaches to the analysis of receiver operating characteristic (ROC) curves," Med. Decis. Making **4**, 137–150 (1984).
[43]B. J. McNeil, E. Keeler, and S. J. Adelstein, "Primer on certain elements of medical decision making," J. Nucl. Med. **17**, 293 (1976).
[44]J. A. Swets and R. M. Pickett, *Evaluation of Diagnostic Systems. Methods from Signal Detection Ttheory* (Academic, New York, 1982).
[45]University of Chicago receiver operating characteristic program software downloads, http://xray.bsd.uchicago.edu/krl/KRL_ROC/software_index6.htm, last checked June 20, 2007.
[46]University of Iowa receiver operating characteristic program software downloads, http://perception.radiology.uiowa.edu/, last checked June 20, 2007.
[47]Free-response receiver operating characteristic software downloads, http://www.devchakraborty.com/downloads.html, last checked June 20, 2007.
[48]H. E. Rockette, W. Li, M. L. Brown, C. A. Britton, J. T. Towers, and D. Gur, "Statistical test to assess rank-order imaging studies," Acad. Radiol. **8**, 24–30 (2001).
[49]W. F. Good *et al.*, "Observer sensitivity to small differences: a multipoint rank order experiment," AJR Am. J. Roentgenol. **173**, 275–278 (1999).
[50]C. A. Britton *et al.*, "Subjective quality assessment of computed radiography hand images," J. Digit. Imaging **9**, 21–24 (1996).
[51]J. D. Towers, J. M. Holbert, C. A. Britton, P. Costello, R. Sciulli, and D. Gur, "Multipoint rank order study methodology: Observer issues," Invest. Radiol. **35**, 125–130 (2000).
[52]D. Gur, D. A. Rubin, B. H. Kart, A. M. Peterson, C. R. Fuhrman, H. E. Rockette, and J. L. King, "Forced choice and ordinal discrete rating assessment of image quality: A comparison," J. Digit. Imaging **10**, 103–107 (1997).
[53]R. M. Slone, D. H. Foos, B. R. Whiting, E. Muka, D. A. Rubin, T. K. Pilgram, K. S. Kohm, S. S. Young, P. Ho, and D. D. Hendrickson, "Assessment of visually lossless irreversible image compression: Comparison of three methods by using an image-comparison workstation," Radiology **240**, 869–877 (2000).
[54]K. H. Lee, Y. H. Kim, B. H. Kim, K. J. Kim, T. J. Kim, H. J. Kim, and S. Hahn, "Irreversible JPEG 2000 compression of abdominal CT for primary

interpretation: Assessment of visually lossless threshold," Eur. Radiol. **17**, 1529–1534 (2007).
[55]R. M. Slone, E. Muka, and T. K. Pilgram, "Irreversible JPEG compression of digital chest radiographs for primary interpretation: Assessment of visually lossless threshold," Radiology **228**, 425–429 (2003).
[56]O. Kocsis, L. Costaridou, L. Varaki, E. Likaki, C. Kalogeropoulou, S. Skiadopoulos, and G. Panayiotakis, "Visually lossless threshold determination for microcalcification detection in wavelet compressed mammograms," Eur. Radiol. **13**, 2390–2396 (2003).
[57]H. Ringl, R. E. Schernthaner, A. A. Bankier, M. Weber, M. Prokop, C. J. Herold, and C. Schaefer-Prokop, "JPEG2000 compression of thin-section CT images of the lung: Effect of compression ratio on image quality," Radiology **240**, 869–877 (2006).
[58]H. S. Woo, K. J. Kim, T. J. Kim, S. Hahn, B. Kim, Y. H. Kim, and K. H. Lee, "JPEG 2000 compression of abdominal CT: Difference in tolerance between thin- and thick-section images," AJR Am. J. Roentgenol. **189**, 535–541 (2007).
[59]C. E. Metz, "Some practical issues of experimental design and data analysis in radiological ROC studies," Invest. Radiol. **24**, 234–245 (1989).
[60]T. Kobayashi, X. W. Xu, H. MacMahon, C. E. Metz, and K. Doi, "Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs," Radiology **199**, 843–848 (1996).
[61]C. E. Metz, in *Handbook of Medical Imaging*, edited by J. Beutel, H. L. Kundel, and R. L. Van-Metter (SPIE, Bellingham, WA, 2000), Vol. 1, pp. 751–769.
[62]S. V. Beiden *et al.*, "Independent versus sequential reading in ROC studies of computer-assist modalities: Analysis of components of variance," Acad. Radiol. **9**, 1036–1043 (2002).
[63]H. E. Rockette, W. L. Campbell, C. A. Britton, J. M. Holbert, J. L. King, and D. Gur, "Empiric assessment of parameters that affect the design of multireader receiver operating characteristic studies," Acad. Radiol. **6**, 723–729 (1999).
[64]N. A. Obuchowski and R. C. Zepp, "Simple steps for improving multiple-reader studies in radiology," AJR Am. J. Roentgenol. **166**, 517–521 (1996).
[65]J. L. King, C. A. Britton, D. Gur, H. E. Rockette, and P. L. Davis, "On the validity of the continuous and discrete confidence rating scales in receiver operating characteristic studies," Invest. Radiol. **28**, 962–963 (1993).
[66]H. E. Rockette, D. Gur, and C. E. Metz, "The use of continuous and discrete confidence judgments in Receiver operating characteristic studies of diagnostic imaging techniques," Invest. Radiol. **27**, 169–172 (1992).
[67]K. S. Berbaum, D. D. Dorfman, E. A. Franken, Jr., and R. T. Caldwell, "An empirical comparison of discrete ratings and subjective probability ratings," Acad. Radiol. **9**, 756–763 (2002).
[68]American College of Radiology (ACR), *The Breast Imaging Reporting and Data System Atlas* (American College of Radiology, Reston, VA, 2004).
[69]W. E. Barlow *et al.*, "Accuracy of screening mammography interpretation by characteristics of radiologists," J. Natl. Cancer Inst. **96**, 1840–1850 (2004).
[70]J. J. Fenton *et al.*, "Influence of computer-aided detection on performance of screening mammography," N. Engl. J. Med. **356**, 1399–1409 (2007).
[71]R. F. Wagner, S. V. Beiden, and C. E. Metz, "Continuous versus categorical data for ROC analysis: Some quantitative considerations," Acad. Radiol. **8**, 328–334 (2001).
[72]Y. Jiang, R. M. Nishikawa, R. A. Schmidt, C. E. Metz, M. L. Giger, and K. Doi, "Improving breast cancer diagnosis with computer-aided diagnosis," Acad. Radiol. **6**, 22–33 (1999).
[73]D. D. Dorfman, K. S. Berbaum, and C. E. Metz, "Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method," Invest. Radiol. **27**, 723–731 (1992).
[74]N. A. Obuchowski, "Multireader, multimodality receiver operating characteristic curve studies: Hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations," Acad. Radiol. **2**, 522–529; Acad. Radiol. **2**, S57–S64; Acad. Radiol. **2**, S70–S21 (1995).
[75]A. Toledano and C. A. Gatsonis, "Regression analysis of correlated receiver operating characteristic data," Acad. Radiol. **2**, S30–S36; Acad. Radiol. **2**, S61–S34; Acad. Radiol. **2**, S70–S31 (1995).
[76]S. L. Hillis, N. A. Obuchowski, K. M. Schartz, and K. S. Berbaum, "A comparison of the Dorfman–Berbaum–Metz and Obuchowski–Rockette methods for receiver operating characteristic (ROC) data," Stat. Med. **24**, 1579–1607 (2005).
[77]S. V. Beiden, R. F. Wagner, and G. Campbell, "Components-of-variance models and multiple-bootstrap experiments: An alternative method for random-effects, receiver operating characteristic analysis," Acad. Radiol. **7**, 341–349 (2000).
[78]S. V. Beiden, R. F. Wagner, G. Campbell, C. E. Metz, and Y. Jiang, "Components-of-variance models for random-effects ROC analysis: The case of unequal variance structures across modalities," Acad. Radiol. **8**, 605–615 (2001).
[79]S. V. Beiden, R. F. Wagner, G. Campbell, and H. P. Chan, "Analysis of uncertainties in estimates of components of variance in multivariate ROC analysis," Acad. Radiol. **8**, 616–622 (2001).
[80]H. H. Barrett, M. A. Kupinski, and E. Clarkson, "Probabilistic foundations of the MRMC method," Proc. SPIE **5749**, 21–31 (2005).
[81]B. D. Gallas, "One-shot estimate of MRMC variance: AUC," Acad. Radiol. **13**, 353–362 (2006).
[82]F. Wang and C. A. Gatsonis, "Hierarchical models for ROC curve summary measures: Design and analysis of multi-reader, multi-modality studies of medical tests," Stat. Med. **27**, 243–256 (2008).
[83]S. J. Starr, C. E. Metz, L. B. Lusted, and D. J. Goodenough, "Visual detection and localization of radiographic images," Radiology **116**, 533–538 (1975).
[84]R. G. Swensson, "Unified measurement of observer performance in detecting and localizing target objects on images," Med. Phys. **23**, 1709–1725 (1996).
[85]P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons, "A free-response approach to the measurement and characterization of radiographic-observer performance," J. Appl. Photogr. Eng. **4**, 166–171 (1978).
[86]D. P. Chakraborty and K. S. Berbaum, "Observer studies involving detection and localization: Modeling, analysis, and validation," Med. Phys. **31**, 2313–2330 (2004).
[87]D. C. Edwards, C. E. Metz, and M. A. Kupinski, "Ideal observers and optimal ROC hypersurfaces in N -class classification," IEEE Trans. Med. Imaging **23**, 891–895 (2004).
[88]X. He, C. E. Metz, B. M. Tsui, J. M. Links, and E. C. Frey, "Three-class ROC analysis—A decision theoretic approach under the ideal observer framework," IEEE Trans. Med. Imaging **25**, 571–581 (2006).
[89]D. P. Chakraborty, "Recent advances in observer performance methodology: Jackknife free-response ROC (JAFROC)," Radiat. Prot. Dosimetry **114**, 26–31 (2005).
[90]D. P. Chakraborty, "Analysis of location specific observer performance data: Validated extensions of the jackknife free-response (JAFROC) method," Acad. Radiol. **13**, 1187–1193 (2006).
[91]B. Zheng, D. P. Chakraborty, H. E. Rockette, G. S. Maitz, and D. Gur, "A comparison of two data analyses from two observer performance studies using jackknife ROC and JAFROC," Med. Phys. **32**, 1031–1034 (2005).
[92]J. Shiraishi, D. Appelbaum, Y. Pu, Q. Li, L. Pesce, and K. Doi, "Usefulness of temporal subtraction images for identification of interval changes in successive whole-body bone scans: JAFROC analysis of radiologists' performance," Acad. Radiol. **14**, 959–966 (2007).
[93]K. Ueda, S. Iwasaki, M. Nagasawa, S. Sueyoshi, J. Takahama, K. Ide, and K. Kichikawa, "Hard-copy versus soft-copy image reading for detection of ureteral stones on abdominal radiography," Radiat. Med. **21**, 210–213 (2003).
[94]E. A. Berns, R. E. Hendrick, M. Solari, L. Barke, D. Reddy, J. Wolfman, L. Segal, P. DeLeon, S. Benjamin, and L. Willis, "Digital and screen-film mammography: comparison of image acquisition and interpretation times," AJR Am. J. Roentgenol. **187**, 38–41 (2006).
[95]H. M. Zafar, R. S. Lewis, and J. H. Sunshine, "Satisfaction of radiologists in the United States: A comparison between 2003 and 1995," Radiology **244**, 223–231 (2007).
[96]A. Zuger, "Dissatisfaction with medical practice," N. Engl. J. Med. **350**, 69–75 (2004).
[97]S. P. Prabhu, S. Gandhi, and P. R. Goddard, "Ergonomics of digital imaging," Br. J. Radiol. **78**, 582–586 (2005).
[98]P.L. Spath, "Caring on empty: Fatigue in healthcare," Radiol. Today, July, 20–24 (2006).
[99]T. Vertinsky and B. Forster, "Prevalence of eye strain among radiologists: Influence of viewing variables on symptoms," AJR Am. J. Roentgenol. **184**, 681–686 (2005).
[100]E. A. Krupinski and M. Kallergi, "Choosing a radiology workstation:

technical and clinical considerations," Radiology **242**, 671–682 (2007).
[101]S. Halligan, D. G. Altman, S. Mallett, S. A. Taylor, D. Burling, M. Roddie, L. Honeyfield, J. McQuillan, H. Amin, and J. Dehmeshki, "Computed tomographic colonography: Assessment of radiologist performance with and without computer-aided detection," Gastroent. **131**, 2006–2009 (2006).
[102]S. Kakeda, K. Kamada, Y. Hatakeyama, T. Aoki, Y. Korogi, S. Katsuragawa, and K. Doi, "Effect of temporal subtraction technique on interpretation time and diagnostic accuracy of chest radiography," AJR Am. J. Roentgenol. **187**, 1253–1259 (2006).
[103]S. H. Kim, J. M. Lee, Y. J. Kim, J. Y. Choi, G. H. Kim, H. Y. Lee, and B. I. Choi, "Detection of hepatocellular carcinoma on CT in liver transplant candidates: Comparison of PACS tile and multisynchronized stack modes," AJR Am. J. Roentgenol. **188**, 1337–1342 (2007).
[104]C. Mariani, A. Tronchi, L. Oncini, O. Pirani, and R. Murri, "Analysis of the x-ray work flow in two diagnostic imaging departments with and without a RIS/PACS system," J. Digit. Imaging **19**, 18–28 (2006).
[105]B. I. Reiner, E. L. Siegel, and K. M. Siddiqui, in *Decision Support in the Digital Medical Enterprise*, edited by B. I. Reiner, E. L. Siegel, and B. J. Erickson (Society for Computer Applications in Radiology, Great Falls, VA, 2005), pp. 121–133.
[106]K. M. Schartz, K. S. Berbaum, R. T. Caldwell, and M. T. Madsen, "Workstation J: Workstation emulation software for medical image perception and technology evaluation research," Proc. SPIE **6515**, 1–11 (2007).
[107]E. A. Krupinski, "Using the human observer to assess medical image display quality," J. Soc. Inf. Disp. **14**, 927–932 (2006).
[108]W. J. Tuddenham and W. P. Calvert, "Visual search patterns in roentgen diagnosis," Radiology **76**, 255–256 (1961).
[109]E. L. Thomas and E. L. Lansdown, "Visual search patterns of radiologists in training," Radiology **81**, 288–291 (1963).
[110]H. L. Kundel, C. F. Nodine, and D. P. Carmody, "Visual scanning, pattern recognition and decision-making in pulmonary tumor detection," Invest. Radiol. **13**, 175–181 (1978).
[111]E. A. Krupinski, "Visual scanning patterns of radiologists searching mammograms," Acad. Radiol. **3**, 137–144 (1996).
[112]C. F. Nodine, C. Mello-Thoms, H. L. Kundel, and S. P. Weinstein, "Time course of perception and decision making during mammographic interpretation," AJR Am. J. Roentgenol. **179**, 917–923 (2002).
[113]E. A. Krupinski, "Technology and perception in the 21st-century reading room," J. Am. Coll. Radiol. **3**, 433–439 (2006).
[114]C. F. Nodine, H. L. Kindel, L. C. Toto, and E. A. Krupinski, "Recording and analyzing eye-position data using a microcomputer workstation," Behav. Res. Methods Instrum. Comput. **24**, 475–485 (1992).
[115]E. Krupinski, H. Roehrig, and T. Furukawa, "Influence of film and monitor display luminance on observer performance and visual search," Acad. Radiol. **6**, 411–418 (1999).
[116]E. A. Krupinski and H. Roehrig, "The influence of a perceptually linearized display on observer performance and visual search," Acad. Radiol. **7**, 8–13 (2000).
[117]E. A. Krupinski, H. Roehrig, J. Fan, and T. Yoneda, "High luminance monochrome vs low luminance monochrome and color softcopy displays: Observer performance and visual search efficiency," Proc. SPIE **65150R**, 105 (2007).
[118]E. A. Krupinski, K. Siddiqui, E. Siegel, R. Shrestha, E. Grant, H. Roehrig, and J. Fan, "Influence of 8-bit vs 11-bit displays on observer performance and visual search: A multi-center evaluation," J. Soc. Inf. Disp. **15**, 385–390 (2007).
[119]E. A. Krupinski and P. J. Lund, "Differences in time to interpretation for evaluation of bone radiographs with monitor and film viewing," Acad. Radiol. **4**, 177–182 (1997).
[120]P. R. G. Bak, "Will the use of irreversible compression become a standard of practice?" SIIM News **18**, 1–10 (2006).
[121]Y. Zhang, B. T. Pham, and M. P. Eckstein, "The effect of nonlinear human visual system components on performance of a channelized Hotelling observer model in structured backgrounds," IEEE Trans. Med. Imaging **25**, 1348–1362 (2006).
[122]Y. Jiang, D. Huo, and D. L. Wilson, "Methods for quantitative image quality evaluation of MRI parallel reconstructions: Detection and perceptual difference model," Magn. Reson. Imaging **25**, 712–721 (2007).
[123]W. B. Jackson, P. Beebee, D. A. Jared, D. K. Biegelsen, J. O. Larimer, J. Lubin, and J. L. Gille, "X-ray system design using a human visual model," Proc. SPIE **2708**, 29–40 (1996).
[124]E. Krupinski, J. Johnson, H. Roehrig, and J. Lubin, "Using a human visual system model to optimize soft-copy mammography display: Influence of display phosphor," Acad. Radiol. **10**, 161–166 (2003).
[125]J. P. Johnson, J. Nafziger, E. A. Krupinski, J. Lubin, and H. Roehrig, "Effects of grayscale window/level parameters on breast-lesion detectability," Proc. SPIE **5034**, 462–473 (2003).
[126]E. A. Krupinski, J. Lubin, H. Roehrig, J. Johnson, and J. Nafziger, "Using a human visual system model to optimize soft-copy mammography display: Influence of veiling glare," Acad. Radiol. **13**, 289–295 (2006).
[127]N. A. Obuchowski, "Sample size tables for receiver operating characteristic studies," AJR Am. J. Roentgenol. **175**, 603–608 (2000).
[128]J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology **143**, 29–36 (1982).
[129]J. A. Hanley and B. J. McNeil, "A method of comparing the areas under receiver operating characteristic curves derived from the same cases," Radiology **148**, 839–843 (1983).
[130]X. H. Zhou, N. A. Obuchowski, and D. K. McClish, *Statistical Methods in Diagnostic Medicine* (Wiley, New York, 2002).
[131]H. P. Chan *et al.*, "Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis," Invest. Radiol. **25**, 1102–1110 (1990).
[132]W. P. Kegelmeyer, J. M. Pruneda, P. D. Bourland, A. Hillis, M. W. Riggs, and M. L. Nipper, "Computer-aided mammographic screening for spiculated lesions," Radiology **191**, 331–337 (1994).
[133]D. Gur, A. I. Bandos, C. R. Fuhrman, A. H. Klym, J. L. King, and H. E. Rockette, "The prevalence effect in a laboratory environment: Changing the confidence ratings," Acad. Radiol. **14**, 49–53 (2007).
[134]K. S. Berbaum, G. Y. El-Khoury, E. A. Franken, D. M. Kuehn, D. M. Meis, D. D. Dorfman, N. G. Warnock, B. H. Thompson, S. C. S. Kao, and M. H. Kathol, "Missed fractures resulting from satisfaction of search effect," Emerg. Radiol. **1**, 242–249 (1994).
[135]K. S. Berbaum, E. A. Franken, D. D. Dorfman, E. M. Miller, E. A. Krupinski, K. Kreinbring, R. T. Caldwell, and C. H. Lu, "The cause of satisfaction of search effects in contrast studies of the abdomen," Acad. Radiol. **3**, 815–826 (1996).
[136]K. S. Berbaum, G. Y. El-Khoury, K. Ohashi, K. M. Schartz, R. T. Caldwell, M. T. Madsen, and E. A. Franken, "Satisfaction of search in multitrauma patients: Severity of detected fractures," Acad. Radiol. **14**, 711–722 (2007).
[137]C. T. Loy and L. Irwig, "Accuracy of diagnostic tests read with and without clinical information: A systematic review," JAMA, J. Am. Med. Assoc. **292**, 1602–1609 (2004).
[138]L. Ruess, S. C. O'Connor, K. H. Cho, F. H. Hussain, W. J. Howard, R. C. Slaughter, and A. Hedge, "Carpal tunnel syndrome and cubital tunnel syndrome: Work-related musculoskeletal disorders in four symptomatic radiologists," AJR Am. J. Roentgenol. **181**, 37–42 (2003).
[139]E. A. Krupinski, A. Johns, and K. S. Berbaum, "Measurement of visual strain in radiologists," Proc. SPIE (in press).
[140]J. M. Lewin *et al.*, "Comparison of full-field digital mammography with screen-film mammography for cancer detection: Results of 4,945 paired examinations," Radiology **218**, 873–880 (2001).
[141]T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center," Radiology **220**, 781–786 (2001).
[142]D. Gur *et al.*, "Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system," J. Natl. Cancer Inst. **96**, 185–190 (2004).
[143]E. D. Pisano *et al.*, "American College of Radiology Imaging Network digital mammographic imaging screening trial: Objectives and methodology," Radiology **236**, 404–412 (2005).
[144]E. D. Pisano *et al.*, "Diagnostic performance of digital versus film mammography for breast-cancer screening," N. Engl. J. Med. **353**, 1773–1783 (2005).
[145]J. Warwick, L. Tabar, B. Vitak, and S.W. Duffy, "Time-dependent effects on survival in breast carcinoma: results of 20 years of follow-up from the Swedish Two-County Study," Cancer **100**, 1331–1336 (2004).
[146]National Lung Screening Trial (NLST) National Cancer Institute web site, http://www.cancer.gov/nlst, last checked August 231, 2007.
[147]P. B. Bach, J. R. Jett, U. Pastorino, M. S. Tockman, S. J. Swensen, and C. B. Begg, "Computed tomography screening and lung cancer outcomes," JAMA, J. Am. Med. Assoc. **297**, 953–961 (2007).
[148]P. M. Marcus, E. J. Bergstralh, M. H. Zweig, A. Harris, K. P. Offord, and R. S. Fontana, "Extended lung cancer incidence follow-up in the Mayo

Lung Project and overdiagnosis," J. Natl. Cancer Inst. **98**, 748–756 (2006).
[149]M. M. Oken, P. M. Marcus, P. Hu, T. M. Beck, W. Hocking, P. A. Kvale, J. Cordes, T. L. Riley, S. D. Winslow, S. Peace, D. L. Levin, P. C. Prorok, and J. K. Gohagan, "Baseline chest radiograph for lung cancer detection in the randomized Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial," J. Natl. Cancer Inst. **97**, 1832–1839 (2005).
[150]J. L. Weissfeld, R. E. Schoen, P. F. Pinsky, R. S. Bresalier, T. Church, S. Yurgalevitch, J. H. Austin, P. C. Prorok, and J. K. Gohagan, "Flexible sigmoidoscopy in the PLCO cancer screening trial: Results from the baseline screening examination of a randomized trial," J. Natl. Cancer Inst. **97**, 989–997 (2005).
[151]G. L. Andriole, D. L. Levin, E. D. Crawford, E. P. Gelmann, P. F. Pinsky, D. Chia, B. S. Kramer, D. Reding, T. R. Church, R. L. Grubb, G. Izmirlian, L. R. Ragard, J. D. Clapp, P. C. Prorok, and J. K. Gohagan, "Prostate Cancer Screening in the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial: Findings from the initial screening round of a randomized trial," J. Natl. Cancer Inst. **97**, 433–438 (2005).
[152]S. S. Buys, E. Partridge, M. H. Greene, P. C. Prorok, D. Reding, T. L. Riley, P. Hartge, R. M. Fagerstrom, L. R. Ragard, D. Chia, G. Izmirlian, M. Fouad, C. C. Johnson, and J. K. Gohagan, "Ovarian cancer screening in the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial: Findings from the initial screen of a randomized trial," Am. J. Obstet. Gynecol. **193**, 1630–1639 (2005).
[153]R. M. Nishikawa, in *Digital Mammography*, edited by S. M. Astley, M. Brady, C. Rose, and R. Zwiggelaar (Springer, London, 2006), pp. 46–53.
[154]N. Breen *et al.*, "Reported drop in mammography: Is this cause for concern?," Cancer **109**, 2405–2409 (2007).
[155]L. Tabar, S. W. Duffy, B. Vitak, H. H. Chen, and T. C. Prevost, "The natural history of breast carcinoma: What have we learned from screening?," Cancer **86**, 449–462 (1999).
[156]D. A. Berry, "Benefits and risks of screening mammography for women in their forties: A statistical appraisal," J. Natl. Cancer Inst. **90**, 1431–1439 (1998).
[157]D. A. Berry *et al.*, "Effect of screening and adjuvant therapy on mortality from breast cancer," N. Engl. J. Med. **353**, 1784–1792 (2005).
[158]P. C. Gotzsche and O. Olsen, "Is screening for breast cancer with mammography justifiable?," Lancet **355**, 129–134 (2000).
[159]O. Olsen and P. C. Gotzsche, "Cochrane review on screening for breast cancer with mammography," Lancet **358**, 1340–1342 (2001).
[160]S. A. Feig, R. L. Birdwell, and M. N. Linver, "Computer-aided screening mammography," N. Engl. J. Med. **357**, 84; author reply, N. Engl. J. Med. **357**, 85 (2007).
[161]A. Jemal, R. Siegel, E. Ward, T. Murray, J. Xu, and M. J. Thun, "Cancer statistics, 2007," Ca Cancer J. Clin. **57**, 43–66 (2007).
[162]Y. Jiang, D. L. Miglioretti, C. E. Metz, and R. A. Schmidt, "Breast cancer detection rate: Designing imaging trials to demonstrate improvements," Radiology **243**, 360–367 (2007).
[163]R. Ballard-Barbash *et al.*, "Breast Cancer Surveillance Consortium: A national mammography screening and outcomes database," AJR Am. J. Roentgenol. **169**, 1001–1008 (1997).
[164]J. G. Elmore, C. K. Wells, C. H. Lee, D. H. Howard, and A. R. Feinstein, "Variability in radiologists' interpretations of mammograms," N. Engl. J. Med. **331**, 1493–1499 (1994).
[165]C. A. Beam, P. M. Layde, and D. C. Sullivan, "Variability in the interpretation of screening mammograms by US radiologists. Findings from a national sample," Arch. Intern. Med. **156**, 209–213 (1996).
[166]D. Gur, "Objectively measuring and comparing performance levels of diagnostic imaging systems and practices," Acad. Radiol. **14**, 641–642 (2007).
[167]C. M. Rutter and S. Taplin, "Assessing mammographers' accuracy. A comparison of clinical and test performance," J. Clin. Epidemiol. **53**, 443–450 (2000).
[168]K. Moberg, N. Bjurstam, B. Wilczek, L. Rostgard, E. Egge, and C. Muren, "Computed assisted detection of interval breast cancers," Eur. J. Radiol. **39**, 104–110 (2001).
[169]C. Marx *et al.*, "Are unnecessary follow-up procedures induced by computer-aided diagnosis (CAD) in mammography? Comparison of mammographic diagnosis with and without use of CAD," Eur. J. Radiol. **51**, 66–72 (2004).
[170]E. Alberdi *et al.*, "Use of computer-aided detection (CAD) tools in screening mammography: A multidisciplinary investigation," Br. J. Radiol. **78**, S31–S40 (2005).
[171]P. Taylor and R. M. Given-Wilson, "Evaluation of computer-aided detection (CAD) devices," Br. J. Radiol. **78**, S26–S30 (2005).
[172]F. J. Gilbert *et al.*, "Single reading with computer-aided detection and double reading of screening mammograms in the United Kingdom National Breast Screening Program," Radiology **241**, 47–53 (2006).
[173]S. H. Taplin, C. M. Rutter, and C. D. Lehman, "Testing the effect of computer-assisted detection on interpretive performance in screening mammography," AJR Am. J. Roentgenol. **187**, 1475–1482 (2006).
[174]G. M. te Brake, N. Karssemeijer, and J. H. Hendriks, "Automated detection of breast carcinomas not detected in a screening program," Radiology **207**, 465–471 (1998).
[175]L. J. Warren Burhenne *et al.*, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," Radiology **215**, 554–562 (2000).
[176]R. L. Birdwell, D. M. Ikeda, K. F. O'Shaughnessy, and E. A. Sickles, "Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection," Radiology **219**, 192–202 (2001).
[177]B. Zheng, R. Shah, L. Wallace, C. Hakim, M. A. Ganott, and D. Gur, "Computer-aided detection in mammography: An assessment of performance on current and prior images," Acad. Radiol. **9**, 1245–1250 (2002).
[178]R. F. Brem *et al.*, "Improvement in sensitivity of screening mammography with computer-aided detection: A multiinstitutional trial," AJR Am. J. Roentgenol. **181**, 687–693 (2003).
[179]N. Karssemeijer *et al.*, "Computer-aided detection versus independent double reading of masses on mammograms," Radiology **227**, 192–200 (2003).
[180]S. V. Destounis, P. DiNitto, W. Logan-Young, E. Bonaccio, M. L. Zuley, and K. M. Willison, "Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience," Radiology **232**, 578–584 (2004).
[181]D. M. Ikeda, R. L. Birdwell, K. F. O'Shaughnessy, E. A. Sickles, and R. J. Brenner, "Computer-aided detection output on 172 subtle findings on normal mammograms previously obtained in women with breast cancer detected at follow-up screening mammography," Radiology **230**, 811–819 (2004).
[182]S. Ciatto *et al.*, "Computer-aided detection (CAD) of cancers detected on double reading by one reader only," Breast **15**, 528–532 (2006).
[183]P. Skaane, A. Kshirsagar, S. Stapleton, K. Young, and R. A. Castellino, "Effect of computer-aided detection on independent double reading of paired screen-film and full-field digital screening mammograms," AJR Am. J. Roentgenol. **188**, 377–384 (2007).
[184]M. C. Difazio *et al.*, "Digital chest radiography: Effect of temporal subtraction images on detection accuracy," Radiology **202**, 447–452 (1997).
[185]L. Monnier-Cholley, H. MacMahon, S. Katsuragawa, J. Morishita, T. Ishida, and K. Doi, "Computer-aided diagnosis for detection of interstitial opacities on chest radiographs," AJR Am. J. Roentgenol. **171**, 1651–1656 (1998).
[186]D. J. Getty, R. M. Pickett, C. J. D'Orsi, and J. A. Swets, "Enhanced interpretation of diagnostic images," Invest. Radiol. **23**, 240–252 (1988).
[187]H. P. Chan *et al.*, "Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: An ROC study," Radiology **212**, 817–827 (1999).
[188]K. Ashizawa *et al.*, "Effect of an artificial neural network on radiologists' performance in the differential diagnosis of interstitial lung disease using chest radiographs," AJR Am. J. Roentgenol. **172**, 1311–1315 (1999).
[189]J. Shiraishi, H. Abe, R. Engelmann, M. Aoyama, H. MacMahon, and K. Doi, "Computer-aided diagnosis to distinguish benign from malignant solitary pulmonary nodules on radiographs: ROC analysis of radiologists' performance–initial experience," Radiology **227**, 469–474 (2003).
[190]M. A. Helvie *et al.*, "Sensitivity of noncommercial computer-aided detection system for mammographic breast cancer detection: Pilot clinical trial," Radiology **231**, 208–214 (2004).
[191]R. L. Birdwell, P. Bandodkar, and D. M. Ikeda, "Computer-aided detection with screening mammography in a university hospital setting," Radiology **236**, 451–457 (2005).
[192]T. E. Cupples, J. E. Cunningham, and J. C. Reynolds, "Impact of computer-aided detection in a regional screening mammography program," AJR Am. J. Roentgenol. **185**, 944–950 (2005).
[193]L. A. Khoo, P. Taylor, and R. M. Given-Wilson, "Computer-aided detection in the United Kingdom National Breast Screening Programme: Prospective study," Radiology **237**, 444–449 (2005).
[194]J. C. Dean and C. C. Ilvento, "Improved cancer detection using computer-

aided detection with diagnostic and screening mammography: Prospective study of 104 cancers," AJR Am. J. Roentgenol. **187**, 20–28 (2006).
[195]J. M. Ko, M. J. Nicholas, J. B. Mendel, and P. J. Slanetz, "Prospective assessment of computer-aided detection in interpretation of screening mammography," AJR Am. J. Roentgenol. **187**, 1483–1491 (2006).
[196]M. J. Morton, D. H. Whaley, K. R. Brandt, and K. K. Amrami, "Screening mammograms: Interpretation with computer-aided detection–prospective evaluation," Radiology **239**, 375–383 (2006).
[197]S. A. Feig, E. A. Sickles, W. P. Evans, and M. N. Linver, "Re: Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system," J. Natl. Cancer Inst. **96**, 1260–1261; author reply, J. Natl. Cancer Inst. **96**, 1261 (2004).
[198]D. Gur, "Computer-aided screening mammography," N. Engl. J. Med. **357**, 83–84; author reply, N. Engl. J. Med. **357**, 85 (2007).
[199]R. M. Nishikawa, R. A. Schmidt, and C. E. Metz, "Computer-aided screening mammography," N. Engl. J. Med. **357**, 84; author reply, N. Engl. J. Med. **357**, 85 (2007).
[200]American College of Radiology (ACR), *The Breast Imaging Reporting and Data System Atlas* (American College of Radiology, Reston, VA, 2004), p. 195.
[201]C. A. Roe and C. E. Metz, "Variance-component modeling in the analysis of receiver operating characteristic index estimates," Acad. Radiol. **4**, 587–600 (1997).