Anniversary Paper: Evaluation of medical imaging systems

Elizabeth A. Krupinski
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 and in telemedicine with digital image acquisition for dermatology, ophthalmology, and cardiology. 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.

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. Key to NIBIB’s mission is “supporting studies to assess the effectiveness and outcomes of new biologics, materials, processes, devices, and procedures.” 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.

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 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.
(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.

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. 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, low-dose CT for lung cancer screening, and CADe and CADx tools.

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.

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. Molecular imaging, emerging with significant potential for measuring the outcome of individualized cancer therapy, is also an area of great interest today.

The highest level of efficacy is on the society as a whole and can be very difficult to measure quantitatively. Evaluation of societal efficacy 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. 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. Radiologists and imaging equipment manufacturers naturally disagree with this view and the cut. 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 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. 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. 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, ROC analysis was developed by researchers from such diverse fields as engineering, psychology, and mathematics. 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.[53,64] 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). Although the BI-RADS final assessment categories 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, recall versus routine screening, etc.—provides
additional information on the reader’s diagnostic decision that is complementary to the ROC curve and provides important information for cost-benefit analyses.

A multitude of statistical methods has been developed to analyze the MRMC experiment, 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, and Dorfman, Berbaum, and Metz developed the first practical algorithm for this analysis. Their method allows meaningful comparison of modalities, simultaneously accounting for both reader-skill variation and case-difficulty variation. Alternative methods have also been developed, and Hillis et al. showed recently a close relationship (or equivalence) between two of these methods.

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. With their method, it is now possible to quantify explicitly the contribution of reader variability in a MRMC experiment. Gallas later developed a different method based on an approach proposed by Barrett et al. that provides similar estimates without invoking the method of bootstrapping. It is likely that new methods for analyzing the MRMC experiment will continue to be developed in the near future.

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, 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. 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 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.

One area of recent development of ROC methods is FROC analysis. 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 and simulations suggest that the JAFROC method may yield greater statistical power than other methods, including the Dorfman–Berbaum–Metz method.

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. In both fields, the transition to digital imaging systems has resulted in significant increases in the amount of time required to view a case and surveys suggest that work overload contributes substantially more to clinician dissatisfaction than in the past.

There is evidence that the electronic reading room leads to greater fatigue and some are investigating whether that impacts diagnostic performance. 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. 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...
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). Similarly, temporal subtraction (13.6 s per case without subtraction versus 10.8 s with subtraction) and stack mode presentation of multi-slice image data (75 s per case for tile mode versus 63 s for stack mode) significantly reduce interpretation time.

Viewing time can be measured simply with a stopwatch or sophisticatedly with computer auditing tools that automatically record every interaction between the radiologist and a workstation. Another sophisticated evaluation method is eye-position recording (see Fig. 2). Eye-tracking studies have been used to gain basic understanding of the visual search and decision-making process and also for system evaluation. 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

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

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.
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.

Higher display luminance, calibration with the Digital Imaging and Communications in Medicine (DICOM) Grayscale Display Function standard, high-performance monochrome rather than color display, and 11-bit rather than 8-bit display 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. Understanding how the GUI affects users accessing information is useful for both GUI designers and radiologists.

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. 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.

A useful approach to system evaluation that does not require human observers is modeling. 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. 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. These models often correlate well with human performance and have been used to evaluate imaging system components such as phosphor for CRT display, image window/level over an entire image versus over a small region covering a lesion, the effect of display veiling glare on performance, and reconstruction algorithms for parallel MRI with multiple coils and k-space subsampling. 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. 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. There are other useful guidelines and tables in the literature, which do not have accompanying computer programs.

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) 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.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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,\textsuperscript{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.\textsuperscript{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\textsuperscript{138} and visual and physical fatigue begin to emerge,\textsuperscript{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.\textsuperscript{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,\textsuperscript{7} it is necessary to distinguish laboratory tests, field tests, and mortality trials. We call the typical laboratory observer study\textsuperscript{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.\textsuperscript{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.\textsuperscript{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. 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. 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? 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 opportunite 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. 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) 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.

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, and (lung 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. 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. 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. 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. 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), 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)—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.

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 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. 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.\textsuperscript{131,132,168–173} There is also a body of literature that shows the potential benefit of CADe in screening mammography\textsuperscript{174–183} and of similar computer aids in other diagnostic tasks.\textsuperscript{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.\textsuperscript{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.\textsuperscript{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,\textsuperscript{162} conclusions were nonetheless drawn suggesting that computer-aided detection is not associated with improved detection of breast cancer.\textsuperscript{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.\textsuperscript{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\textsuperscript{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.\textsuperscript{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\textsuperscript{68} and these scales have been used as a basis for ROC analysis.\textsuperscript{69,101} 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.\textsuperscript{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.\textsuperscript{3} However, in principle,\textsuperscript{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\textsuperscript{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 systems 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.

\textsuperscript{1}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


96. E. A. Kupinski and M. Kallergi, “Choosing a radiology workstation:


