HHS Public Access

Author manuscript

Acad Radiol. Author manuscript; available in PMC 2017 August 15.

Published in final edited form as:

Acad Radiol. 2012 April ; 19(4): 463–477. doi:10.1016/j.acra.2011.12.016.

Evaluating Imaging and Computer-aided Detection and Diagnosis Devices at the FDA

Brandon D. Gallas, PhD, **Heang-Ping Chan, PhD**, **Carl J. D'Orsi, MD**, **Lori E. Dodd, PhD**, **Maryellen L. Giger, PhD**, **David Gur, ScD**, **Elizabeth A. Krupinski, PhD**, **Charles E. Metz, PhD**, **Kyle J. Myers, PhD**, **Nancy A. Obuchowski, PhD**, **Berkman Sahiner, PhD**, **Alicia Y. Toledano, ScD**, and **Margarita L. Zuley, MD**

Division of Imaging and Applied Mathematics, Center for Devices and Radiological Health, US Food and Drug Administration, 10903 New Hampshire Avenue, Building 62, Room 3124, Silver Spring, MD 20993-0002 (B.D.G., K.J.M., B.S.); the Department of Radiology, University of Michigan, Ann Arbor, Michigan (H.-P.C.); Emory University, Atlanta, Georgia (C.J.D.); Biostatistics Research Branch, National Institute of Allergy and Infectious Diseases, Bethesda, Maryland (L.E.D.); the Department of Radiology, University of Chicago, Chicago, Illinois (M.L.G., C.E.M.); the Department of Radiology, University of Pittsburgh School of Medicine (D.G.); the Department of Radiology, University of Arizona, Tucson, Arizona (E.A.K.); Quantitative Health Sciences, Cleveland Clinic Foundation, Cleveland, Ohio (N.A.O.); Statistics Collaborative, Inc, Washington District of Columbia (A.Y.T.); and the Department of Radiology, University of Pittsburgh (M.L.Z.)

Abstract

This report summarizes the Joint FDA-MIPS Workshop on Methods for the Evaluation of Imaging and Computer-Assist Devices. The purpose of the workshop was to gather information on the current state of the science and facilitate consensus development on statistical methods and study designs for the evaluation of imaging devices to support US Food and Drug Administration submissions. Additionally, participants expected to identify gaps in knowledge and unmet needs that should be addressed in future research. This summary is intended to document the topics that were discussed at the meeting and disseminate the lessons that have been learned through past studies of imaging and computer-aided detection and diagnosis device performance.

Keywords

Reader studies; designs; methods; ROC; premarket; postmarket; consensus

In this report, we summarize the Joint FDA-MIPS Workshop on Methods for the Evaluation of Imaging and Computer-Assist Devices, held on July 14, 2010. The purpose of this workshop was to gather information and facilitate consensus on the current state of the science of statistical methods and study designs for the evaluation of imaging devices and computer-aided detection and diagnosis (CAD) systems in the hands of clinicians, the image readers, to support US Food and Drug Administration (FDA) submissions. We hope that this

Address correspondence to: B.D.G. brandon.gallas@fda.hhs.gov.

summary serves as a stand-alone starting point and overview for investigators who are interested in reader studies evaluating imaging technology to support FDA submissions.

Additionally, we identified gaps in knowledge and unmet needs that should be addressed in future research. This document summarizes the topics that were discussed and the lessons learned from past studies of readers using imaging devices and CAD systems. Our primary goal is to improve device evaluations submitted to the FDA, that is, to make them more meaningful and powerful, thereby reducing the resources required of the FDA and industry to make important and clinically useful devices and tools available for improved patient care.

The workshop invitations targeted a multidisciplinary group of leaders in the field of medical imaging evaluation who have somehow been involved in the FDA approval and clearance of imaging devices: statisticians, clinicians, imaging physicists, mathematicians, and computer scientists from government, academia, and industry. All told, there were 22 attendees from academia, 11 from industry, 31 from the FDA, five from other government agencies including the National Institutes of Health, and 12 clinicians. Most attendees had either attended or participated in an FDA advisory panel meeting to judge the safety and effectiveness of an imaging device seeking premarket approval.

The meeting was divided into three sessions, each with two invited presenters followed by a panel that included an additional invited expert in the field. The first session was “Statistical Perspectives,” with Alicia Toledano, ScD (Statistics Collaborative, Inc), and Nancy Obuchowski, PhD (Cleveland Clinic Foundation), as the key presenters and Berkman Sahiner, PhD (FDA), on the panel. The second session was “Clinical Perspectives,” with Carl D’Orsi, MD (Emory University), and Margarita Zuley, MD (University of Pittsburgh), as key presenters and Barbara McNeil, MD, PhD (Harvard Medical School), as the third member on the panel. The final session was “Developer Perspectives,” with Heang-Ping Chan, PhD (University of Michigan), and Maryellen Giger, PhD (University of Chicago), as key presenters and David Gur, ScD (University of Pittsburgh), on the panel. The sessions were moderated by Elizabeth Krupinski, PhD (University of Arizona), and Lori Dodd, PhD (National Institute of Allergy and Infectious Diseases), who also provided a summary titled “State of Consensus and Plans for the Future” at the end of the meeting. Kyle Myers, PhD, and Brandon Gallas, PhD, of the FDA organized and provided the introductory background and guidance for the workshop.

Because the main focus of the workshop was the FDA approval and clearance of imaging devices, the views expressed there may have been limited or biased. The discussions at this workshop were focused on getting imaging devices approved and did not define what is or is not a valuable contribution to the scientific literature. Given the high stakes, manufacturers are sometimes reluctant to use cutting-edge study designs and analyses, preferring more established and traditional designs and analyses. In this setting, manufacturers typically report sensitivity and specificity, receiver operating characteristic (ROC) curves and areas under ROC curves (AUCs). However, they are not sure which of these limited measures is the “right” one (the one that will get their devices approved) and which measure they should use to size their studies. Manufacturers also often struggle to balance study biases against

study burden (complexity, time, and costs) and clinical reality against experimental tractability and abstraction.

The background section of this summary introduces diagnostic decisions and performance to the uninitiated. Readers familiar with sensitivity, specificity, their relationship through the ROC curve, thresholds, optimal thresholds (cost-benefit analysis), and the AUC summary measure may want to only skim this subsection. This background section does not discuss the entire spectrum of evaluation methods, though location-based image evaluation methods are treated later in "Topics Deserving More Research and Attention." The background section does discuss the role of human readers (the clinicians interpreting the images) and reader variability. Without reader variability, many of the challenges in study design and analysis would not exist.

The background section also outlines the phases of evaluation, which helps define the setting for studies that support FDA submissions. Without this setting, the comments and recommendations might appear to overreach, because studies to support FDA submission occur at a very specific and singular point in the life cycle of a technology or device. This point marks the technology's transition from its premarket life to its postmarket life. One of the biases of the workshop might be that the clinical context and the clinical work flow for an imaging device are more or less unknown before marketing and that the final niche that a new device will occupy in the clinical arena will evolve after marketing over time. This bias reduces the emphasis on clinical reality in favor of experimental tractability and abstraction.

Following the background material, this summary outlines consensus recommendations on topics that determine the balance between study biases and study burden, clinical reality, and experimental tractability. Then, before concluding, we identify topics that deserve more research and attention because they have the potential to improve the evaluation of imaging devices.

Development of this summary used the speaker talks and related discussions, summaries from the two moderators, and a follow-up survey of the workshop speakers on key potential consensus statements. The goal of the survey was to get explicit opinions from all speakers on each statement, as well as to elicit discussion and references to the literature.

BACKGROUND

The context for the workshop was the evaluation of medical imaging devices with reader studies to support FDA decision making, often using CAD as an example. All the criteria and procedures by which the FDA makes its determinations can be found in the Code of Federal Regulations, Title 21, Part 860 (1). The criteria relevant to the workshop and this summary are codified in the definition of effectiveness: "There is reasonable assurance that a device is effective when it can be determined, based upon valid scientific evidence, that in a significant portion of the target population, the use of the device for its intended uses and conditions of use, when accompanied by adequate directions for use and warnings against unsafe use, will provide clinically significant results."

The FDA's definition of effectiveness is precise and crisp as it relates to the target population and device use, and it is broad and flexible as it relates to "reasonable assurance" and "clinically significant results." The focus of the discussions reported here was on the broad and flexible parts of the definition, namely, what studies and analyses have heretofore been successful in bringing to market medical imaging devices that have a screening or diagnostic indication for use (IFU). "IFU" is an FDA term defined as "a general description of the disease or condition the device will diagnose, treat, prevent, cure, or mitigate, including a description of the patient population for which the device is intended" (2).

This focus was chosen because imaging devices, CAD products, image analysis packages, and display devices represent a large and growing portion of the Center for Devices and Radiological Health regulatory portfolio. Also, a Center for Devices and Radiological Health panel meeting in March 2008 had emphasized the need for new paradigms for the evaluation of these products using reader studies and standardized databases (3). These evaluations must be statistically interpretable, relevant for their IFUs, and at a reasonable cost.

The focus did not include IFUs related to image-guided surgery or biopsy or the use of imaging to evaluate the effectiveness of therapies (eg, drugs, surgery, other interventions). The use of images for quantification or measurement (eg, reporting lesion size, ejection fraction), although part of some CAD systems, also was not considered in the workshop's limited focus.

Diagnostic Decisions and Performance

It is straightforward to evaluate the diagnostic performance of an imaging device when the task involves a binary decision. Does an image show disease or not? A binary classification task is often modeled as a two-step process. In the first step, "the reader" ranks patients on a scale corresponding to the reader's belief or confidence that each patient has the disease in question. In the second step of the model, "the reader" decides whether this rating (often referred to as a likelihood, probability, or confidence that the disease is present) is high enough to classify the patient as diseased, that is, by comparing the rating to a threshold. In the case of algorithmic (computer or model) readers, this two-step process is often explicit; a measurement or score is produced and then compared to a threshold. With human readers, these two steps are not explicit and perhaps do not model the human decision process. However, even for a human reader, a data collection method can be used that first obtains the reader's basic binary decision and then asks the reader to provide a rating of his or her confidence. Such a method allows the investigator to include or remove the threshold from the performance evaluation, as well as to consider performance with other thresholds. This additional information comes at little to no cost in time and does not distort or add noise to the binary decision.

It is true that clinicians do not typically rate their confidence in clinical practice; instead, they decide upon or recommend the next clinical action. However, ratings are not unprecedented. For example, Breast Imaging Reporting and Data System (BI-RADS) ratings (4) divide category 4 into three subdivisions: 4A, 4B, and 4C. This subdivision "allows a more meaningful practice audit, is useful in research involving receiver-operating

characteristic (ROC) curve analysis, and is an aid for clinicians and pathologists" (4). It is utility in research, or more specifically here, utility in imaging technology evaluation, that motivates the collection of ratings in addition to a binary decision, as we describe below.

For any particular threshold, there is a corresponding pair of performance measures: sensitivity and specificity. For example, in screening mammography, the sensitivity (the true-positive fraction) is typically given as the proportion of women with cancer who are correctly recalled for additional imaging, and specificity (the true-negative fraction) is the proportion of women without cancer who are correctly not recalled for additional imaging. A reader's operating point is the particular sensitivity-specificity pair at which he or she performs.

An ROC curve illustrates the trade-off between the sensitivity and specificity of the reader across all thresholds (5). This trade-off is realized by a change in the reader's threshold setting. In the case of breast cancer screening via mammography, when the threshold is made more aggressive, the reader recalls more patients for additional imaging, increasing his or her sensitivity at the price of lower specificity. If the reader's threshold is moved in the opposite direction, the reader will recall fewer patients; the reader is less aggressive, decreasing his or her sensitivity with the concomitant result of increased specificity.

The optimal trade-off between sensitivity and specificity can be considered for a given ROC curve, disease prevalence, and relative weights of the risks and benefits of the four possible decision and truth (actual disease status) combinations: true and false positives, true and false negatives. Determination of the optimal decision strategy is referred to as utility analysis (also referred to as cost-benefit, or risk-benefit, analysis) (6–11). Although this determination of the optimal trade-off is tractable in principle, it is complicated in practice. First, the utility weights vary on the basis of many variables introduced by differences in the patient, the reader, and the health care environment that may include subjective considerations. Next, the utility weights can change over time for many reasons, such as the cost and effectiveness of subsequent tests, treatments, and improved knowledge of all the above. Moreover, the trade-off that is optimal in clinical practice for a mature technology may not translate to reader behavior in the early investigation of reader performance for a new technology. We return to this point later in this summary.

Although the individual operating points and the shape of the ROC curve are important, in many instances, one may seek a statistic that summarizes the ROC curve in a single index. A common summary statistic is AUC, the area under the ROC curve (12). One interpretation of AUC is that it is a reader's average sensitivity over all possible specificities. The AUC measures how well a reader is able to separate the population of diseased patients from nondiseased patients. As such, it is a global summary of the ROC curve that avoids thresholds entirely.

When the AUC of one test is higher than another, the reader has the potential to operate at a higher sensitivity for any specificity or a higher specificity for any sensitivity. Care must be taken when comparing two modalities' AUCs. Good practice dictates that investigators actually look at the ROC plots when calculating AUCs. They may find that ROC curves

substantially cross or that there are questionable interpolations or extrapolations of the measured operating points (13–15). In these cases, the AUC is not an appropriate summary of the ROC curve. Additionally, for specific applications, the AUC may include portions of the ROC space that are not of practical use (eg, regions of low specificity). Therefore, in some applications, it may be more appropriate to restrict analysis to a more clinically relevant region (See “Sensitivity for a ‘Clinically Relevant’ Specificity or Specificity Range”).

In practice, choosing a multilevel rating scale for readers to report during an ROC study is not trivial. Overall, the workshop speakers were uncomfortable with rating levels that were calibrated or somehow related to clinical actions, the reason being that clinical actions are not typically ordinal and are usually few in number, such that they rarely cover a substantial range of ROC space. BI-RADS ratings (4), for example, are problematic for ROC studies, because they are not strictly ordinal and they sample a very limited region in ROC space (16). The workshop speakers were more comfortable with a generic, multilevel (possibly 101-point) scale, on which “absolutely sure no disease is present” is associated with the lowest level and “absolutely sure disease is present” is associated with the highest level. Such a generic scale has the potential to sample ROC space more adequately and can be augmented (at little cost to the investigator) by explicitly linking, or anchoring, clinical actions to different points on the generic scale (see “Linking Binary and Multilevel Ratings”).

The paradigm of ROC analysis, and the measurement of the AUC in particular, is essential to the field of diagnostic imaging assessment. This is especially true in radiology, in which the science of clinical decision making and ROC analysis has had a long and productive history, resulting in a great many tutorials, literature, and other practical discussions related to conducting ROC experiments and the analysis of their results (13,17–19).

Reader Variability

When estimating the performance of an imaging device, it is also essential to estimate the uncertainty in the performance estimate. The uncertainty in an estimate such as AUC determined from a reader study includes contributions from the cases in the study as well as the readers. Reader variability occurs because human readers are often not reproducible (internally or across readers), especially in terms of scoring and classification of patient images. Human readers can have both a wide range of diagnostic ability (skill) as well as variability in their levels of aggressiveness (operating point). Demonstrations of this reader variability in the interpretation of mammograms are given in reports by Elmore et al (20) and Beam et al (21). One solution to overcoming some of this variability is to plot the results in ROC space, as recommended by D’Orsi et al (22). In ROC space, one can see a predictable relationship between sensitivity and specificity. This relationship is the reason for ROC curves. The ROC curve eliminates the variability in the readers’ different operating points and allows one to concentrate on diagnostic ability and its variability.

Reader variability severely challenges the evaluation and comparison of imaging devices and acquisition protocols. To overcome this challenge, investigators now recognize that studies need multiple readers and that an efficient study design is one in which each reader reads

each case in all conditions being evaluated (eg, the new technology vs the "standard of care"). This study design is often referred to as the fully crossed multireader, multicase (MRMC) reader study. However, there are many other MRMC study designs, for example, studies in which different readers read different cases and studies that allow for variations in the number of cases read by each reader (23–25). Although these not fully crossed MRMC study designs complicate statistical analysis and often reduce statistical power, they may also be deemed practical or efficient in terms of sharing cases across sites, controlling the amount of reads required by a single reader, or using modality expertise.

Tools and software are available in the literature and online (26–28) to size and analyze MRMC studies. As such, MRMC studies are now expected in situations beyond exploratory studies, in which the analysis accounts for reader variability. This sentiment has been reiterated by the editors of *Radiology* (29), who stated in a recent editorial that the audience of their journal was interested in the variability between observers and techniques and that reporting these "requires a sufficiently high number of observers." Likewise, the American College of Radiology Imaging Network (ACRIN) recognizes that there is a great deal of variability among readers and that reader variability "must be accounted for in the design of ACRIN trials" (30).

All methods for sizing and analyzing an MRMC study require understanding the core components of variance and correlation: reader variability, case variability, and the correlation of reader scores within and across the imaging modalities (31). Furthermore, all of these effects depend on the application (ie, disease, technology, reader training, and experience). Consequently, estimating the magnitudes of these components for an MRMC study is best accomplished using relevant preliminary data (usually a pilot study) and appropriate statistical modeling of reader performance and reader variability (32–34). Without this effort, it is impossible to predict meaningfully how many readers and cases will be enough to provide adequate statistical power. Although the exact number of readers for a desired power depends on the effect size and components of variance, the workshop speakers agreed that the importance of reader variability must not be ignored and that ≥10 readers may be needed to reduce the impact of reader variability on statistical power to an acceptable level for a study to support an FDA submission.

Phases of Evaluation

There are many phases in the evaluation of an imaging device that include image interpretation by clinicians. Such evaluations can be grouped into three phases (35): (1) exploratory (early, pilot), (2) challenge (stress, lab based), and (3) advanced (late, clinical use).

Exploratory (early, pilot) studies demonstrate the clinical and engineering proof of concept. These studies might not include any patients or readers but might use simulations, phantoms, excised tissue, and organs. When the studies do involve patients, sample sizes are small (10–50), and they are often conveniently acquired rather than representative of an intended population (eg, the diseased patients may have atypically large lesions). When these exploratory studies involve readers, there are only a few (one to three), reader variability is not evaluated, and they read the images retrospectively. These reader studies serve the

purpose of learning how to use the new technology—how to interpret the resulting images—and how the images can change clinical decisions. The readers and the investigators answer the following questions: What does disease look like? What does normal look like? What can be measured? and How might patients benefit? These early studies help determine the IFU of a device, as well as appropriate end points for the next study phases.

Challenge studies often compare a new technology to the current practice. The goal is not to evaluate clinical performance on an absolute scale but to rank one system against another. The term “challenge” indicates a common practice of using challenging cases and controls (ie, “stressing” the new technology). At the same time, challenge studies often include key subgroups either known or suspected to be handled by the new device differently such that the new technology may outperform or underperform the old for these study populations (36). The number of cases in a challenge study is moderate (hundreds of patients), such that a rigorous subgroup analysis of small differences is beyond reach, but trends and gross differences across subgroups can still be identified, thereby leading to new challenge studies.

Challenge studies for imaging devices typically use archived, or retrospective, images. In large part, this is driven by the practice of selecting challenging cases from specific subgroups (ie, enriching the image data set). The use of retrospective images also provides the opportunity for multiple readers to read the same cases. As mentioned, this helps reduce the confounding effect of reader variability without removing the reader from the evaluation process. Using retrospective images also removes the impact of the study on patient care.

The main differences between a challenge study and an advanced clinical-use study are the purpose and setting. An advanced clinical-use study collects clinical decisions and clinical outcomes in a real-world setting (ie, prospectively). The study conditions in general (and the difficulty of its cases in particular) represent all aspects of the diagnostic setting of interest as accurately as possible. The resulting performance estimates are then on an absolute scale.

An advanced clinical-use study is much more demanding than a challenge study. In contrast to a challenge study, it is designed such that cases are included prospectively, the device may influence patient management, and readers are not necessarily blinded to patient information (eg, chart, history, other tests and prior images). Demands on the number of cases and readers can increase substantially over those of challenge studies, especially in situations in which the disease prevalence is low (37) or important subgroups are underrepresented. The effort required to obtain the standard for truth can be higher. Because these demands are so high, advanced clinical-use studies might also take the form of meta-analyses (37–39). The data analysis may also require complicated decision-analysis models to account for the entire patient management scenario, as is done in cost-benefit and comparative effectiveness studies (40–42).

Digital breast tomosynthesis (DBT) provides an example of a technology that is in the challenge phase of evaluation. One might say that the earlier work by Niklason et al (43) in 1997 on DBT occurred near the end of the exploratory phase. Their study instructed radiologists to read images of phantoms and mastectomy specimens and provide a subjective

five-point score of lesion visibility, lesion margin visibility, and confidence in classifying a lesion as benign or malignant.

The recent evaluation of DBT by Gur et al (44) fits in the challenge phase. This lab-based retrospective study included eight readers who interpreted 125 exams under four display modes. The term “lab based” here means that a study is controlled much more strictly than what would happen in clinical practice. Gur et al’s DBT study differed from clinical practice in the following ways: (1) no prior images or clinical information were provided to the readers, (2) the study population was enriched with cancers and had an overrepresentation of challenging normal cases (noncancers from BI-RADS category 0), (3) multiple readers interpreted each case, and (4) the interpretations did not affect patient management.

The Digital Mammographic Imaging Screening Trial is an example of an advanced clinical-use study (45). Patients were accrued prospectively and managed by the results of full-field digital mammography (FFDM) as well as with the results of standard-practice screen-film mammography (SFM). Readers had access to patient information (ie, prior studies, demographic information, and histories). The study enrolled 49,528 women and 153 radiologists. Unlike most advanced clinical-use studies, the Digital Mammographic Imaging Screening Trial collected seven-point and 101-point ROC scores for each woman in addition to the clinical action BI-RADS scores. The additional reporting scales complemented the clinical action scale. The multiple reporting scales allowed the study to ascertain how radiologists used the image information to manage patients (BI-RADS), as well as how well radiologists were able to separate women with cancer from women without cancer overall (AUC).

PREMARKET VERSUS POSTMARKET

The workshop speakers acknowledged that the future clinical use of an imaging device will likely differ from how it was used in the trial submitted to the FDA to support approval. The final niche that a new device will occupy in the clinical arena (how it fits in with patient management, what patients it works best for) will evolve over time. This evolution is a result of dissemination of the device: more readers using the device on more patients. The dissemination brings with it feedback between the users and the developers, sharing of experiences on small and large scales (clinical rounds, conference presentations, journal papers), and, most important, additional challenge studies. As these data are accumulated, the community gains a better appreciation of performance in the clinical setting. If the technology flourishes, the community of users starts to outline guidelines and best practices, as well as develop training opportunities and the technology comes into widespread use.

The workshop speakers also acknowledged that lab-based challenge studies are not perfect predictors of clinical use for many reasons. First, challenge studies suffer from spectrum bias: the challenge study population does not match the intended-use population because it is enriched with challenging cases (see discussion below). Moreover, challenge studies are biased because the reading conditions are substantially different from clinical reading. For example, blinding a reader to patient information withholds information he or she typically uses, quality assurance and quality control of image acquisition and display may be different

from those in the real world, and using checklists causes the reader to take more time and be more deliberate. Also, the mind-set and behavior of the reader can be affected by prevalence (46,47), by knowing that his or her decision will not affect a patient and simply by knowing that he or she is being studied (ie, the “Hawthorne effect” (48,49)). The sum total of these effects has been referred to as the “laboratory effect” and was investigated recently by Gur et al (50) for mammographic interpretations.

Despite the biases in a lab-based challenge study, and because the clinical niche of an imaging device evolves over time, the workshop speakers generally agreed that FDA approvals of novel imaging devices should be based on all the evaluations (concept, physics, biology, bench tests) leading up to and culminating in a comparative lab-based challenge study. In addition to labeling that clearly describes the device evaluations, comparing the new technology to a reference on an identical or similar set of cases can mitigate the concerns regarding the biases, enabling a lab-based challenge study to provide a reasonable assurance of effectiveness to put the device on the market. An adequately sized prospective clinical-use study can be too burdensome when the expected prevalence and detection rate are low.

Comfort with lab-based challenge studies has been reinforced by the studies conducted to bring FFDM on the market. A review of the summaries of safety and effectiveness shows that the clinical data that helped the four original FFDM submissions (by GE, Fischer, Hologic, and Fuji (51–54)) gain FDA approval were not collected in a prospective real-world clinical environment:

- The readers evaluating the images were blinded to the patients’ prior studies, demographic information, and history.
- The study sizes were moderate (200–600 cases), and the case sampling was enriched to include more cancers (50–125 cancers). The enrichment targeted challenging cases, using the results of SFM for case selection (women with abnormal findings on SFM, women undergoing diagnostic mammography because of SFM, or women recommended for biopsy via SFM).
- The studies had between five and 12 readers reading every case (fully crossed) in each modality (readers and cases were paired across modalities).
- The interpretations collected clinical action reports, but the primary end point in all was noninferiority (FFDM vs SFM) on the basis of AUCs (ROC scores were probabilities of malignancy or levels of suspicion; three were on a 101-point scale, and one was on a five-point scale).

Each of the studies showed that the AUCs of the full-field digital mammographic devices were within a noninferiority margin of 0.10 to those of the screen-film mammographic comparators. When the comparison was made in the Digital Mammographic Imaging Screening Trial, an advanced clinical-use trial, noninferiority was validated with much more precision.

Although the workshop speakers recommended lab-based challenge studies for imaging devices, they also allowed exceptions that would motivate a more advanced clinical-use

study. If there is evidence that a lab-based reader study may give meaningfully different results from the ultimate clinical use, evaluations might need to be based on actual clinical use and observed clinical outcomes. Alternatively, there can be a compromise, whereby a lab-based challenge study is adequate for getting the device on the market conditional on an FDA-ordered postmarket study.

The workshop speakers did not have a unanimous opinion about whether more FDA-ordered postmarket studies were needed. However, the group did feel that this need increases with the magnitude of (direct or indirect) potential risks to the patient when the device is used. Ideally, there should be consolidation and dissemination of the results and outcomes of all studies that happen after a device is on the market. The obvious problem is organizing and paying for such an effort.

In this digital world, though, more information can be recorded and tracked with automation. For example, it should not be complicated for CAD software to output a report on the locations and scores it calculates and save these reports in an electronic file (especially in mammography, in which there is already a systematic tracking effort with the Mammography Quality Standards Act). It is important that device manufacturers make such tools available to further study the performance of devices after FDA approval. Some workshop participants stressed FDA-industry collaboration or an independent entity for encouraging the design and use of such tools. Such tracking would be enormously useful for comparing clinical-use results to lab-based studies submitted for FDA approval, monitoring the evolution of CAD software updates, and uncovering adverse effects from “off-label” use (using a CAD system in a manner different from the IFU). Besides giving the community the information to improve public health, the information can provide feedback for the regulatory process, for example, by providing data for subgroups that may not have been possible to adequately investigate in lab-based studies and by informing the FDA on what type of updates could be cleared with minimal oversight (because they are minor or can be expected to yield positive results) and what type of updates should require more careful scrutiny. At the end of the day, it takes the community to see benefit in collecting this kind of data, to find a way to convince manufacturers to make such monitoring possible, and to organize and pay for the data collection and analysis.

AREA UNDER THE CURVE VERSUS SENSITIVITY AND SPECIFICITY

The workshop speakers agreed that the AUC is usually a reasonable and useful measure of effectiveness for diagnostic imaging in many situations, especially for FDA submissions. Some in the group indicated that the AUC was the most reasonable and useful measure, while others wanted to leave room for exceptions (eg, when ROC curves cross and when humans are unable to provide more than a binary response). The reason the AUC is particularly useful is because it removes the variability and ambiguity of a reader’s level of aggressiveness (19,21,22,34,36,55), which is the manifestation of the reader’s internal risk-benefit trade-off between calling a patient diseased or not. When the AUC is appropriate, the workshop speakers felt the study needs to be designed and sized only to show the desired effect in AUC with appropriate confidence and power. The study does not need to be designed and sized also to show an effect in either sensitivity or specificity.

The study designs appropriate for a comparative lab-based reader study differ from those for a clinical-use study, and some aspects of these differences affect the reader-perceived a priori risks for the patient and the risks for making decisions. Differences of this kind may influence readers' decision behavior, making it difficult to generalize sensitivity and specificity estimates from a lab-based reader study to clinical performance. Even so, the group believed that it is useful to collect sensitivity and specificity as secondary end points to monitor and appreciate the ultimate role of the technology in clinical practice. The information gained can be applied later to appropriately train, or calibrate, readers. Ultimately, sensitivity and specificity are best characterized in advanced clinical-use studies. Likewise, functions of sensitivity and specificity (eg, likelihood ratios and predictive values) also are best left to advanced clinical-use studies in the postmarket setting.

The statistical argument for preferring the AUC as opposed to sensitivity and specificity is clear: the AUC is less variable. The AUC retains all of the information from the ordered response scale, which inherently has more information than the binary scale. The AUC uses information from all the cases at the same time. For MRMC studies, the statistical argument is even stronger, because all measures are affected by the variability in the skill of readers, but sensitivity and specificity are additionally affected by variability in the level of aggressiveness. Beam et al's (21) data are an excellent demonstration of these variability differences.

READER STUDY DESIGNS AND ANALYSES

The workshop speakers brought up several specific topics related to study designs and analyses for evaluating imaging devices. Some of these topics are about common sense, some are best practices, and others identify the leading edge: study designs and analyses that promise more precision with fewer resources and results that are more relevant to the clinical task.

In a given study, several end points may be worthy of investigation. However, each additional end point can substantially increase the size of the study. The workshop speakers believed that the number of primary end points needs to be limited for sizing a clinical trial.

Enrichment

Matching the spectrum of a study population to a screening population is impractical in most instances. For one reason, it takes too many cases without disease to get enough cases with disease. Equally important, in the clinic, many cases are obvious; they will not stress, or challenge, the ability (or difference in ability) of any imaging device or protocol. These cases add to the burden of conducting an evaluation study without increasing statistical power for demonstrating differences in the ability of competing imaging modalities to classify cases. There are several case-sampling methods that are designed to increase the power of comparison in reader studies or equivalently reduce the size. We broadly refer to these sampling methods as enrichment: increasing prevalence and/or case difficulty. Enrichment methods, which trade the unbiased absolute performance results for the practical ability to compare imaging devices with possible moderate biases, are often acceptable.

Some enrichment is simply the result of including cases with and without disease in a proportion different from that seen in clinical practice. Other enrichment is driven by some measurement or evaluation, for example, a prescreening panel may remove large obvious lesions, or the report by the enrollment clinician or other test results may be used to increase the likelihood of including a case with disease. Considering the direction and magnitude of biases introduced by these enrichment methods should be a part of any conscientiously designed study.

In the case of simply changing the prevalence, there are no mathematical biases, absolute or relative, when estimating sensitivity, specificity, or AUC. However, biases due to changed human behavior may remain. According to utility analysis, a reader should be more aggressive in a higher prevalence setting; the reader should trade off more false-positives for the same or better true positives. This behavior has been observed in studies of radiologic devices (34,47,56,57), making it difficult to compare imaging devices and generalize lab-based estimates of sensitivity and specificity to the clinical environment. In contrast, Gur et al (46) found that no significant effect to the AUC could be measured as a function of prevalence (from 2% to 28%) in a laboratory environment.

When the enrichment method is based on a measurement (other imaging results, diagnostic tests, patient characteristics, or risk factors), the case spectrum is distorted, biasing sensitivity, specificity, and AUC. The challenge is that the nature of these mathematical biases is not well understood. The direction of the mathematical biases was discussed in a “candid assessment” by Lewin (58) for the case in which the enrichment method is based on one of the imaging devices in the comparison (SFM vs FFDM). He pointed out that the “use of a positive SFM exam for enrollment biases the results toward SFM in terms of sensitivity for cancer detection and toward FFDM in terms of specificity.” The results of a study by Cole et al (59) are consistent with these biases.

More recently, some progress has been made to understand enrichment bias (60), recognizing the relationship to survey sampling and verification bias (61–65). What has been learned is that the direction and magnitude of enrichment bias depends on the correlation between the measurement used to drive the enrichment (eg, screen-film mammographic interpretations at enrollment) and the ROC ratings collected from the imaging devices being compared (eg, screen-film and full-field digital mammographic interpretations in the reader study) (60). The higher the correlation, the higher the bias in sensitivity and specificity. As such, it is better to enrich using measurements that are independent of the imaging modalities that are being compared or are based on measurements from both. For example, in comparing SFM to FFDM, it would be better to enrich using positive results on SFM and positive results on FFDM, instead of just SFM.

Recent work (60) mathematically validated the bias claim by Lewin (58) and provided mathematical tools to correct for that bias. Unfortunately, correcting the bias can be accompanied with a substantial increase in variance (60). Regarding the AUC, the picture is still unclear. The specific direction and magnitude are a complicated function of the enrichment method and the correlations involved (60). As such, future work needs to

investigate different enrichment methods and determine the correlations that actually occur before recommendations on correcting for the biases can be made.

Pairing Cases and Readers Across Modalities

When comparing performance across modalities, there are statistical advantages to using the same cases in both modalities. Some refer to this concept as pairing cases across modalities or using a case as its own control. The statistical advantage is obvious to statisticians but bears repeating. In imaging, the advantage extends to pairing the readers across modalities.

The advantage of pairing cases comes from the correlation between the ROC ratings of different imaging modalities for a given case. Pairing cases attempts to control the case difficulty and other confounding factors seen by the two modalities. The degree to which a lesion is visible in one modality often correlates with its visibility in another. This correlation reduces variability in measuring differences. The variance of a difference between A and B is given by

$$\mathrm{Var}(\mathrm{A}-\mathrm{B})=\mathrm{Var}(\mathrm{A})+\mathrm{Var}(\mathrm{B})-2\mathrm{Corr}(\mathrm{A},\mathrm{B})\sqrt{\mathrm{Var}(\mathrm{A})\mathrm{Var}(\mathrm{B})}.$$

The higher the correlation, the lower the variance of the difference.

The advantage of pairing readers is due to the correlation of skill across modalities of an individual reader and his or her scoring tendencies. The correlations related to the reader impact the variability in the same way as the correlations related to the case.

A fully crossed study design takes this concept one step further. In this study design, every reader reads every case in both modalities, building in more correlations that lead to additional reductions in variances of performance differences. This study design is quite popular and has several analysis methods designed for it (31,66–70).

The workshop speakers agreed that a fully crossed study design is the preferred study design in most instances. The additional statistical power achieved by building in correlations can have an enormous impact on the size of a study. The main concerns related to this study design are based on the risks of imaging the patient twice and the feasibility of collecting the additional images and readings. For example, it may not be appropriate to expose a patient to ionizing radiation more than once. In real-time imaging procedures (colonoscopy, colposcopy), it is not feasible to bring in multiple readers to perform the procedure.

In addition to the statistical power that is gained by pairing readers and cases, or by crossing all readers with all cases, the practical implications are even larger. The cost of enrolling cases is doubled in the unpaired study. The overhead and training of readers is doubled. The fully crossed paired-reader, paired-case study design is becoming known as being the most efficient use of readers and truth-validated cases (71). However, if the condition being detected is not rare, or the bottom-line efficiency of a study also involves the duration of the study or the total number of readings, a mixed design might be appropriate (23–25).

Subgroups

It is generally desirable to know if and how device performance differs across patient subgroups: age, ethnicity, risk factors, disease types, and other confounding abnormalities. For imaging devices, questions also arise regarding reader subgroups. However, powering a study for each subgroup increases the overall sample size more or less in proportion with the number of subgroups and their sizes, because every subgroup is essentially treated as an independent question to be answered by the study. The workshop speakers all felt that studies supporting FDA submissions generally do not need to be sized to show effects in patient or reader subgroups. Of course, if different diseases that the device targets have largely different consequences in terms of risk or treatment, if substantial presentation differences between modalities are expected for specific subgroups (eg, through a known or suspected biologic or physical mechanism), or if the training and experience of reader subgroups are expected to play a key role in using the imaging device, the study should be sized to demonstrate differences across the subgroups. In these situations, the differences in performance for the critical subgroups should be evaluated and identified in the device IFU.

Determining what subgroups are critical is subjective. This causes uncertainty for manufacturers on what is expected from them. The workshop speakers indicated that the community would appreciate guidance on which subgroups are important for each technology or disease and a public forum might be very useful for determining this guidance.

Rather than sizing a study for subgroups that are not deemed critical, the workshop speakers felt that it was good practice to have the different subgroups represented in the study population. An underpowered subgroup analysis of the resulting data could then be viewed as a pilot study for determining whether a focused postmarket study would be appropriate. The results of such a postmarket study might enhance the labeling of a device and would certainly be welcome information for the user community. These comments do not preclude the possibility that, in some cases, unplanned subgroup analyses may reveal the need for more premarket study or a restriction in the label for a device.

Sequential Study Design Versus Independent or Crossover Study Design

With the introduction of computer-aided detection (CADe) devices, a one-arm sequential study design has been introduced that is intended to mimic the IFU of some CADe devices. The CADe IFU referred to here is a sequential one: after the reader makes his or her standard-practice read, the CADe marks are displayed to indicate additional suspicious locations for the reader to consider. This sequential study design collects the findings (ROC scores) made by the clinician before CADe marks are displayed (the unaided reader) and again after (the CADe-aided reader). The performance estimates from unaided and aided readings are then compared. Sequential reading may be used in other applications besides CADe. For example, the studies that supported the approval of the first DBT device, to be used as an adjunct to mammography, also used sequential reading: mammograms were sequentially followed by DBT (44,72).

An alternative to the sequential study design is the independent or crossover study design, which includes two distinct reading sessions. One reading session is a conventional reading without the CADe device (ie, unaided reader). The other reading session (separated in time from the conventional reading to render it “independent”) collects the reader reports for the same cases using CADe in a mode consistent with its IFU (ie, aided reader). The independent design mitigates the potential for bias in the sequential design due to anticipating the coming CADe marks. That CADe anticipation bias may cause the reader to be less vigilant, biasing relative CADe performance positively, or instead the user may compete with the CADe (supervigilant), thereby biasing relative CADe performance negatively.

The workshop speakers agreed that the sequential study design was acceptable for three basic reasons beyond representing intended use. First, the literature has several examples comparing the results of sequential and independent study designs, none of which shows a significant bias (ie, a bias that changes the overall study results) (73–76). Second, the sequential design builds in a correlation between the unaided and aided performance, which improves the power over the independent design to detect a difference in performance. Third, the sequential design is less burdensome compared to the independent design on readers and reading time because all the data are collected in one session.

Some modifications to the sequential study design were mentioned that might mitigate the bias from the reader anticipating the second condition (ie, the CADe marks or the adjunct modality, if sequential reading does not represent intended use). One suggestion was to randomize between the presentation of the second-read condition and no second read. Another was to present random CADe marks at different levels of performance. Unfortunately, none of these modifications has been investigated. The workshop speakers agreed that further research in this area is needed, and that for the time being, pilot studies may investigate the bias.

Lesion Localization

The goal of many clinical imaging tasks is to find a “lesion,” the broad definition of which may include an arterial blockage or some other visually evident, localized, manifestation of disease. More than one lesion may be present in some imaging tasks. These tasks give rise to a variety of ways to collect data, as well as ways to define true-positives and false-positives.

The traditional definitions of true-positives and false-positives consider the patient as the sampling unit. The data collection scores the patient without indicating a location, and the analysis proceeds to estimate a per patient sensitivity and specificity, or ROC curve.

It is becoming more and more common, especially in research studies, to collect location information and use it in the analysis. This data collection, in fact, often mimics what is done in practice: locations are reported. The simplest approach to using location information would be to collect marks (and scores) at the location of each finding and then calculate a per lesion sensitivity; the lesion is the sampling unit, and it is deemed a true-positive if a reader’s mark points to it. Specificity can be defined per patient (any patient without disease

that has no mark), or it can be quantified by how many marks do not point to lesions normalized per patient, a false-positive rate.

Recent work has shown what many believe to be true: there are differences between per patient and per lesion sensitivity and specificity on an absolute scale, especially as the number of lesions per patient increases (77). This is to be expected, as they have different sampling units. On a relative scale, however, the per patient and per lesion analyses were found to give the same qualitative results by Obuchowski et al (77). In two clinical examples, they found that when the sensitivity or specificity was better under the per patient analysis, it was also better under the per lesion analysis (77). They also found that when there is more than one lesion of interest per patient, per lesion analysis has higher statistical power even when the correlation between lesions within the same patient is as high as 0.75. This is what one would also intuitively expect, because a per lesion analysis has more sampling units and thus more statistical power. More studies are desirable to provide further evidence of the generalizability of this result.

ROC methodology has also been adapted to account for location, creating several variants of the data collection and analysis methods: free-response ROC (78,79), alternative free-response ROC (80), localization-response ROC (81,82), clustered ROC data analysis (83), jackknife analysis of free-response ROC (84,85), and initial-detection-and-candidate-analysis of free-response ROC (86). Although most of these methods are “per lesion” analyses (the sampling unity for diseased is a lesion), the choice of the “not diseased” sampling unit differs: most are per case, but some are per region. Under certain modeling assumptions, the theoretical underpinnings show a direct relationship between AUCs from standard and localization-response ROC curves (81,82). As with the location-specific sensitivity and specificity, the location-specific ROC analyses have been shown to give the same qualitative results (they rank imaging devices and tasks in the same order), and they provide greater statistical power than ROC analyses (84,87–89). On the other hand, most or all location-specific data analysis methods proposed to date require additional modeling assumptions, the validity of which may be open to question.

Some discussion at the workshop focused on whether location information is important in the screening stage, because the goal of screening is to identify patients for follow-up, not to determine definitive locations. The response by the workshop speakers was that location is important at screening. If a lesion is found in mammographic screening, the patient goes on to diagnostic evaluation of the perceived lesion location. In computed tomographic colonography, the reader may look at the whole colon with optical colonoscopy after a positive result on computed tomographic colonography, and the locations identified during computed tomographic colonography will guide optical colonoscopy.

The workshop speakers agreed that using location information should be required in defining sensitivity and specificity, and there is a growing comfort in the location-specific ROC methods, even though their use is still rare in FDA submissions. The main limitation has to do with the ability to define truth of the locations. For example, in prostate cancer detection, the true location of disease is very difficult to obtain and correlate with reader

marks. If it is not possible to determine the truth of the locations marked, it is not possible to do a location-specific analysis.

TOPICS DESERVING MORE RESEARCH AND ATTENTION

A number of topics came up during the workshop that are worth sharing with the larger community. The topics are worthy of more research and attention and have the potential to improve the evaluation of imaging devices. However, they are not topics that elicited consensus opinions.

Pilot Studies and Reader Training

A pilot study is an important first step in evaluating a technology to plan a pivotal study. One purpose of a pilot study is to get an estimate of the effect size and estimates of the variance components (readers and cases). These parameters help size the study correctly in terms of the number of cases and the number of readers (26–28). More practically, a pilot study can also reveal pitfalls of a reader study design. A pilot study can uncover basic work flow issues and incomplete or uninformative case report forms. A pilot study can reveal poor imaging protocols, poor reading protocols, and poor reader training. For an example of a pilot study informing a pivotal study, see the work of Gur et al (44) and Good et al (90). The workshop speakers generally believed that not enough resources are put into pilot studies.

With regard to reader training, the consensus was that it is a critical element of any reader study. Reader training needs to be rigorous for both the new and the comparison technology, and it should also train the reader in the data collection methods (eg, navigation of images, definitions of disease types, disease or abnormality types not included in the study, how to annotate the images and record findings, the meaning and implication of binary decisions, using and interpreting the ROC scale). Reader training reduces bias as well as inter-reader and intrareader variability.

This desire for more training is not limited to the needs of the study (91,92). There is not enough reader training for new imaging techniques outside of the technology evaluations. There needs to be practical training for the interpretation of images from new techniques, more guidance in the user manual, and quality assurance monitoring of performance in postmarket use.

In terms of reader operating points, it was generally agreed that it is hard for a reader to change his or her operating point in practice once it is established with training and experience. However, early in their experiences with a device, readers will have flexible operating points and will seek to find their perceived optimal operating points. One demonstration of this flexibility and a good list of related references is the work by Dean and Ilvento (57). The introductory period of experiences with a device is the time when feedback and training will be most effective (93). Feedback is critical if one is trying to change an established threshold or establish a new one.

Sensitivity for a "Clinically Relevant" Specificity or Specificity Range

Partial AUCs have the potential to focus the diagnostic assessment to a region of operating points expected to be relevant to the clinical application (75,94–99). A partial AUC averages sensitivity only over a limited range of specificity or averages specificity over a limited range of sensitivity: the "clinically relevant" range. This diagnostic assessment can be focused even more by considering the sensitivity at a "clinically meaningful" specificity or the specificity at a "clinically meaningful" sensitivity (100). In these cases, the evaluation experiment collects multilevel ROC data (not binary decisions) while focusing the analysis to the desired range or point.

In addition to an appealing "clinical relevance," it is possible that these focused performance measures have more statistical power than the AUC. However, this has yet to be demonstrated in simulation or explored extensively in actual reader studies. The field needs more examples demonstrating how to choose meaningful specificity or sensitivity ranges or points (drawing from the established or expected patient management paradigm), more examples examining the statistical power of these analyses (accounting for the variability in reader skill and their operating point), and more statistical tools and software for executing them.

Linking Binary and Multilevel Ratings

If one desires to collect both clinical action data and ROC ratings, one option would embrace what Bob Wagner referred to as the "best of both worlds" (36): a study in which the reader reports the clinical action or decision and an ROC score at the same time. Many breast imaging studies have collected ROC ratings and clinical action decision for the same case independently (44,45,75,101–103). It can be argued that because clinical action decisions are linked to BI-RADS assessment categories, which are calibrated to malignancy risk, such a parallel data collection method minimizes the risk for inconsistency (eg, a "disease-absent" clinical decision case having a higher ROC score than a "disease-present" case). However, unless the two scales are explicitly linked, inconsistencies cannot be fully eliminated (104,105). In addition, for imaging studies for which the interpretation is less standardized, independently collecting ROC ratings and clinical action decisions may lead to a large number of inconsistencies.

Two methods can be used to explicitly link ROC ratings and clinical action decisions. One method uses a two-step data collection method. In the first step, the reader reports a clinical action or decision (binary, or a few ordinal levels). In the next step, the reader reports a generic, multilevel rating within that action level. For example, consider a clinical task with the following three decisions: "negative," "recommend follow-up with additional testing," and "recommend therapy." Given that the decision for a case is "recommend follow-up with additional testing," the rating for that case could be split into three levels. It could indicate whether, among other "follow-up with additional testing" cases, the case in question is closer to a "negative" decision, is closer to "recommend therapy," or is right in the middle. A second method labels the clinical action points on a multilevel scale and explicitly alerts readers as to what their ROC rating means in terms of clinical action. The workshop speakers indicated that more research is needed to investigate whether such a data collection

method introduces any bias and that careful implementation and reader training are essential. In addition, they noted that this type of data collection is possible only if the clinical decisions can be ordered and are not overlapping with respect to ROC ratings.

Surrogate End Points

The typical end point for a study uses biopsy or outcome data to determine truth. Surrogate end points are substitutes. An example of the use of a surrogate end point is the image interpretation by an expert or an expert panel.

There is little doubt that surrogate end points could greatly reduce the burden in studies evaluating imaging devices. Surrogate end points reduce the cost of follow-up, reduce the challenge of low prevalence, or both. The shortcomings of surrogate end points are that they do not follow a patient to a clinical outcome, are subject to appreciable variability that needs to be addressed (106,107), and may be highly disease and/or technology specific. Several of the workshop speakers indicated that this topic should get more attention. The bottom line is that more studies are needed to demonstrate that surrogate end points are correlated with the clinical outcome in particular settings in which their use is proposed.

Bench Testing and Simulated Images and Lesions

It was agreed that in general, it is not appropriate to use only bench tests and engineering measurements (spatial resolution, temporal resolution, dynamic range, contrast, noise, etc) to approve imaging devices with fundamentally new operating principles or that require new user skills. In particular, qualified readers must learn how to use and interpret images made with the new technology. However, as the technology matures, the community may gain experience with the bench tests and their relationship to the interpretation ability of readers. Once this experience is gained, usually only after an imaging device is on the market, bench tests may be sufficient for approving new imaging devices with similar operating principles.

Likewise, the use of simulated images and lesions for the main reader studies in the FDA approval process was not supported. Such methods are essential to the engineering phases of imaging devices, but there needs to be much more convincing data for simulated images or lesions to be substituted for real clinical images in the evaluation of novel imaging technologies.

CAD Changes

As has happened at the FDA panel meetings discussing guidance for CAD devices in radiology, the workshop speakers did not have any recommendations for stratifying CAD changes as “major” versus “minor” (108,109), However, concern was expressed that improvements in CAD systems may not be incorporated into clinical care, because of delays in FDA submissions and approval for incremental changes. Those voicing opinions indicated that stand-alone performance evaluations are often adequate (comparing a new version of a CAD device to a previous version on the same data set, in terms of both overall performance and consistency) and that any actual differences in the performance of the CAD devices are likely to be overwhelmed by reader and other sources of variability. One approach to reducing the requirements for clinical data on an updated CAD device may be to model the

impact on reader performance with CAD changes, given stand-alone performance and reader study results with a previous CAD version and new stand-alone results (110). Additionally, sequestered databases for examining the performance of incremental changes to CAD systems could support the FDA approval process (see “Data Sets”).

A perhaps more fundamentally important issue raised is that there is not adequate communication between CAD device companies and the users when changes are made. If the users are not aware of the kind of changes and the impact of those changes on the stand-alone performance characteristics, they may not be able to take advantage of the improved performance of the CAD device by adjusting their confidence in the CAD marks accordingly. Whether changes should be accompanied by any additional user training would depend on the specific changes made.

Data Sets

As individuals, we all know how valuable data are, and as a community, we are finally building the culture and infrastructure for sharing these data. The workshop speakers recommended that data accessibility and quality should continue to improve. Data sharing should include not only the images and the clinical outcomes but also the clinical reports and annotations (from the primary caregiver and any quality control or reference readers) generated from the images as well. Data sharing is essential to address many of the remaining open questions related to the evaluation of imaging and CAD technologies.

Some participants saw value in sequestering data sets for evaluating similar CAD devices and tools. The results of these evaluations could support FDA submissions. The reason for sequestering data sets rather than open sharing of data sets (the National Institutes of Health model) is to avoid the bias that results when models are trained and tested on the same data. These sequestered data sets could be held and managed by an independent organization or group dedicated to technology assessment and could be continually reviewed and renewed to incorporate new images. With a sufficiently large sequestered independent database, cases could be randomly selected to match a case distribution required in the testing of a particular system and intended use. If only the AUC results are given to the particular manufacturer and to the FDA, as opposed to performance on individual cases, the integrity of the sequestered database could be maintained for subsequent tests.

There is no rule against using public databases to evaluate an imaging device to support an FDA submission. The FDA has certain requirements and concerns that need to be satisfied and are more stringent than what is required for National Institutes of Health research grants. Key issues among these are patient consent, traceability, and training and testing hygiene. The idea of patient consent should not be foreign to any investigator in medicine and public health. Traceability is an FDA requirement that is more demanding than most investigators know; it is the ability to trace the data (eg, a particular image or report) in a submission all the way back to an actual patient if need be. This poses problems for most public databases that anonymize patient data. Training and testing hygiene is a concept that should be well known to CAD developers. Testing a CAD algorithm on the same data that it has been trained on leads to optimistic estimates of performance and will not be appropriate to support an FDA submission (111–114).

CONCLUSIONS

Collaboration among industry, academia, and the FDA is essential for improving study designs and statistical analyses for FDA submissions involving imaging devices. Assessment of these devices presents unique challenges that can best be approached through a consensus development approach by all parties involved. Laboratory-based reader studies, as opposed to prospective clinical studies, are often the only practical assessment option, especially when the expected disease prevalence is low. This workshop identified and discussed a number of critical components of laboratory-based reader studies, including reader variability, assessment at a specified cutoff (sensitivity-specificity pair) versus the ROC paradigm, enrichment, sequential versus independent reading, lesion localization, data sets, pilot studies, and reader training.

By bringing together experts from industry, academia, and the government to discuss key issues in the assessment of diagnostic imaging devices, the workshop with the follow-up survey was able to identify important areas of consensus and current issues in assessment, gather up-to-date information on these issues, and identify gaps in knowledge that should be addressed in future research. We look forward to repeating this workshop over time to discuss developments on these exciting issues and to confront new issues related to the evaluation of imaging and CAD devices. As the field of technology and practice assessment continues to evolve and more information (methods and data) becomes available, especially as related to the consistency or lack of consistency between premarket approval studies and actual findings in field trials and epidemiologic studies, the consensus presented here may change.

References

1. Determination of safety and effectiveness. 21 CFR § 860 (2011).
2. Premarket approval application. 21 CFR § 814 (2011).
3. US Food and Drug Administration. [Accessed November 17, 2011] Radiological Devices Panel meeting: computer-aided detection devices. Available at: http://www.fda.gov/ohrms/dockets/ac/08/minutes/2008%134349m1.pdf
4. American College of Radiology. Breast Imaging Reporting and Data System (BI-RADS). Reston, VA: American College of Radiology; 2003.
5. Metz CE. ROC methodology in radiologic imaging. Invest Radiol. 1986; 21:720–733. [PubMed: 3095258]
6. Green, DM., Swets, JA. Signal detection theory and psychophysics. New York: John Wiley; 1966.
7. Lusted, LB. Introduction to medical decision making. Springfield, IL: Charles C. Thomas; 1968.
8. McNeil BJ, Keller E, Adelstein SJ. Primer on certain elements of medical decision making. N Engl J Med. 1975; 293:211–215. [PubMed: 806804]
9. Metz CE. Basic principles of ROC analysis. Semin Nucl Med. 1978; 8:283–298. [PubMed: 112681]
10. Patton DD. Introduction to clinical decision making. Semin Nucl Med. 1978; 8:273–282. [PubMed: 754285]
11. Weinstein, MC., Fineberg, HV. Clinical decision analysis. Philadelphia, PA: Saunders; 1980.
12. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982; 143:29–36. [PubMed: 7063747]
13. Metz CE. Some practical issues of experimental design and data analysis in radiological ROC studies. Invest Radiol. 1989; 24:234–245. [PubMed: 2753640]

14. Gur D, Bandos AI, Rockette HE. Comparing areas under receiver operating characteristic curves: potential impact of the “last” experimentally measured operating point. Radiology. 2008; 247:12–15. [PubMed: 18258813]
15. Pesce LL, Metz CE, Berbaum KS. On the convexity of ROC curves estimated from radiological test results. Acad Radiol. 2010; 17:960–968. [PubMed: 20599155]
16. Jiang Y, Metz CE. BI-RADS data should not be used to estimate ROC curves. Radiology. 2010; 256:29–31. [PubMed: 20574083]
17. Metz, CE. Fundamental ROC analysis. In: Beutel, J.Kundel, HL., Van Metter, RL., editors. Handbook of medical imaging, vol. 1: physics and psychophysics. Bellingham, WA: SPIE Press; 2000. p. 751-769.
18. Obuchowski NA. ROC analysis. AJR Am J Roentgenol. 2005; 184:364–372. [PubMed: 15671347]
19. International Commission of Radiation Units and Measurements. ICRU Report 79. Bethesda, MD: International Commission of Radiation Units and Measurements; 2008. Receiver operating characteristic analysis in medical imaging.
20. Elmore JG, Wells CK, Lee CH, et al. Variability in radiologists’ interpretations of mammograms. N Engl J Med. 1994; 331:1493–1499. [PubMed: 7969300]
21. Beam C, Layde PM, Sullivan DC. Variability in the interpretation of screening mammograms by US radiologists. Arch Intern Med. 1996; 156:209–213. [PubMed: 8546556]
22. D’Orsi CJ, Swets JA. Variability in the interpretation of mammograms. N Engl J Med. 1995; 332:1172.
23. Gallas BD, Pennello GA, Myers KJ. Multi-reader multi-case variance analysis for binary data. J Opt Soc Am A. 2007; 24:B70–B80.
24. Gallas BD, Brown DG. Reader studies for validation of CAD systems. Neural Networks. 2008; 21:387–397. [PubMed: 18215501]
25. Obuchowski NA. Reducing the number of reader interpretations in MRMC studies. Acad Radiol. 2009; 16:209–217. [PubMed: 19124107]
26. University of Chicago. [Accessed November 17, 2011] Metz ROC software. Available at: http://metz-roc.uchicago.edu
27. University of Iowa Medical Image Perception Laboratory. [Accessed November 17, 2011] Home page. Available at: http://perception.radiology.uiowa.edu
28. Cleveland Clinic, Quantitative Health Sciences. [Accessed November 17, 2011] Research activities: ROC analysis. Available at: http://www.bio.ri.ccf.org/html/rocanalysis.html
29. Bankier AA, Levine D, Halpern EF, et al. Consensus interpretation in imaging research: is there a better way? Radiology. 2010; 257:14–17. [PubMed: 20851935]
30. Hillman BJ. ACRIN—lessons learned in conducting multi-center trials of imaging and cancer. Cancer Imaging. 2005; 5:S97–S101. [PubMed: 16361142]
31. Roe CA, Metz CE. Variance-component modeling in the analysis of receiver operating characteristic (ROC) index estimates. Acad Radiol. 1997; 4:587–600. [PubMed: 9261459]
32. Rockette HE, Campbell WL, Britton CA, et al. Empiric assessment of parameters that affect the design of multireader receiver operating characteristic studies. Acad Radiol. 1999; 6:723–729. [PubMed: 10887893]
33. Wagner RF, Beiden SV, Campbell G, et al. Assessment of medical imaging and computer-assist systems: lessons from recent experience. Acad Radiol. 2002; 9:1264–1277. [PubMed: 12449359]
34. Wagner RF, Beam CA, Beiden SV. Reader variability in mammography and its implications for expected utility over the population of readers and cases. Med Decis Making. 2004; 24:561–572. [PubMed: 15534338]
35. Zhou, X-H., Obuchowski, NA., McClish, DK. Statistical methods in diagnostic medicine. Hoboken, NJ: John Wiley; 2002.
36. Wagner RF, Metz CE, Campbell G. Assessment of medical imaging systems and computer aids: a tutorial review. Acad Radiol. 2007; 14:723–748. [PubMed: 17502262]
37. Jiang Y, Miglioretti DL, Metz CE, et al. Breast cancer detection rate: designing imaging trials to demonstrate improvements. Radiology. 2007; 243:360–367. [PubMed: 17456866]

38. Barlow WE, Chi C, Carney PA, et al. Accuracy of screening mammography interpretation by characteristics of radiologists. J Natl Cancer Inst. 2004; 96:1840–1850. [PubMed: 15601640]
39. Taplin SH, Rutter CM, Lehman CD. Testing the effect of computer-assisted detection on interpretive performance in screening mammography. AJR Am J Roentgenol. 2006; 187:1475–1482. [PubMed: 17114540]
40. Otero HJ, Rybicki FJ, Greenberg D, et al. Cost-effective diagnostic cardiovascular imaging: when does it provide good value for the money? Int J Cardiovasc Imaging. 2010; 26:605–612. [PubMed: 20446040]
41. Tosteson ANA, Stout NK, Fryback DG, et al. Cost-effectiveness of digital mammography breast cancer screening. Ann Intern Med. 2008; 148:1–10. [PubMed: 18166758]
42. Shaw LJ, Wilson PWF, Hachamovitch R, et al. Improved near-term coronary artery disease risk classification with gated stress myocardial perfusion SPECT. JACC Cardiovasc Imaging. 2010; 3:1139–1148. [PubMed: 21071002]
43. Niklason LT, Christian BT, Niklason LE, et al. Digital tomosynthesis in breast imaging. Radiology. 1997; 205:399–406. [PubMed: 9356620]
44. Gur D, Abrams GS, Chough DM, et al. Digital breast tomosynthesis: observer performance study. AJR Am J Roentgenol. 2009; 193:586–591.45. [PubMed: 19620460]
45. Pisano ED, Gatsonis C, Hendrick E, et al. Diagnostic performance of digital versus film mammography for breast-cancer screening. N Engl J Med. 2005; 353:1773–1783. [PubMed: 16169887]
46. Gur D, Rockette HE, Armfield DR, et al. Prevalence effect in a laboratory environment. Radiology. 2003; 228:10–14. [PubMed: 12832568]
47. Gur D, Bandos AI, Fuhrman CR, et al. The prevalence effect in a laboratory environment: changing the confidence ratings. Acad Radiol. 2007; 14:49–53. [PubMed: 17178365]
48. Franke RH, Kaul JD. The Hawthorne experiments: first statistical interpretation. Am Sociol Rev. 1978; 43:623–643.
49. McCarney R, Warner J, Iliffe S, et al. The Hawthorne effect: a randomised, controlled trial. BMC Med Res Methodol. 2007; 7:30. [PubMed: 17608932]
50. Gur D, Bandos AI, Cohen CS, et al. The “laboratory” effect: comparing radiologists’ performance and variability during prospective clinical and laboratory mammography interpretations. Radiology. 2008; 249:47–53. [PubMed: 18682584]
51. US Food and Drug Administration. [Accessed November 17, 2011] Summary of safety and effectiveness data: GE FFDM (P990066). Available at: http://www.accessdata.fda.gov/cdrh_docs/pdf/P990066b.pdf
52. US Food and Drug Administration. [Accessed November 17, 2011] Summary of safety and effectiveness data: Fischer FFDM (P010017). Available at: http://www.accessdata.fda.gov/cdrh_docs/pdf/P010017b.pdf
53. US Food and Drug Administration. [Accessed November 17, 2011] Summary of safety and effectiveness data: Hologic FFDM (P010025). Available at: http://www.accessdata.fda.gov/cdrh_docs/pdf/P010025b.pdf
54. US Food and Drug Administration. [Accessed November 17, 2011] Summary of safety and effectiveness data: Fuji FFDM (P050014). Available at: http://www.accessdata.fda.gov/cdrh_docs/pdf5/P050014b.pdf
55. Samuelson F, Gallas BD, Myers KJ, et al. The importance of ROC data. Acad Radiol. 2010; 18:257–258.
56. Kundel HL, Revesz G, Stauffer HM. Evaluation of a television image processing system. Invest Radiol. 1968; 3:44–50. [PubMed: 5637128]
57. Dean JC, Ilvento CC. Improved cancer detection using computer-aided detection with diagnostic and screening mammography: prospective study of 104 cancers. AJR Am J Roentgenol. 2006; 187:20–28. [PubMed: 16794150]
58. Lewin JM. Full-field digital mammography. A candid assessment. Diagn Imaging. 1999; 21:40–45.
59. Cole E, Pisano ED, Brown M, et al. Diagnostic accuracy of Fischer Senoscan digital mammography versus screen-film mammography in a diagnostic mammography population. Acad Radiol. 2004; 11:879–886. [PubMed: 15288038]

60. Pinsky P, Gallas BD. Enriched designs for assessing predictive performance—analysis of bias and variance. Stat Med. In press.
61. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. J Am Stat. 1952; 47:663–685.
62. Katz J, Zeger SL. Estimation of design effects in cluster surveys. Ann Epidemiol. 1994; 4:295–301. [PubMed: 7921319]
63. Zhou XH. Correcting for verification bias in studies of a diagnostic test's accuracy. Stat Methods Med Res. 1998; 7:337–353. [PubMed: 9871951]
64. Alonzo TA, Pepe MS. Assessing accuracy of a continuous screening test in the presence of verification bias. J Roy Stat Soc C-App. 2005; 54:173–190.
65. He H, Lyness JM, McDermott MP. Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias. Stat Med. 2009; 28:361–376. [PubMed: 18680124]
66. Swets, JA., Pickett, RM. Evaluation of diagnostic systems: methods from signal detection theory. New York: Academic Press; 1982.
67. Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis: generalization to the population of readers and patients with the jackknife method. Invest Radiol. 1992; 27:723–731. [PubMed: 1399456]
68. Roe CA, Metz CE. Dorfman-Berbaum-Metz method for statistical analysis of multireader, multimodality receiver operating characteristic (ROC) data: validation with computer simulation. Acad Radiol. 1997; 4:298–303. [PubMed: 9110028]
69. Obuchowski NA, Beiden SV, Berbaum KS, et al. Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods. Acad Radiol. 2004; 11:980–995. [PubMed: 15350579]
70. Gallas BD. One-shot estimate of MRMC variance: AUC. Acad Radiol. 2006; 13:353–362. [PubMed: 16488848]
71. Obuchowski NA. Multireader receiver operating characteristic studies: a comparison of study designs. Acad Radiol. 1995; 2:709–716. [PubMed: 9419629]
72. US Food and Drug Administration. [Accessed November 17, 2011] Summary of safety and effectiveness data: Hologic DBT (P080003). Available at: http://www.accessdata.fda.gov/cdrh_docs/pdf8/P080003b.pdf
73. Kobayashi T, Xu XW, MacMahon H, et al. Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. Radiology. 1996; 199:843–848. [PubMed: 8638015]
74. Beiden SV, Wagner RF, Doi K, et al. Independent versus sequential reading in ROC studies of computer-assist modalities: analysis of components of variance. Acad Radiol. 2002; 9:1036–1043. [PubMed: 12238545]
75. Hadjiiski L, Chan H-P, Sahiner B, et al. Improvement in radiologists' characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: an ROC study. Radiology. 2004; 233:255–265. [PubMed: 15317954]
76. Obuchowski NA, Meziane M, Dachman AH, et al. What's the control in studies measuring the effect of computer-aided detection (CAD) on observer performance? Acad Radiol. 2010; 17:761–767. [PubMed: 20457419]
77. Obuchowski NA, Mazzone PJ, Dachman AH. Bias, underestimation of risk, and loss of statistical power in patient-level analyses of lesion detection. Eur Radiol. 2010; 20:584–594. [PubMed: 19763582]
78. Bunch PC, Hamilton JF, Sanderson GK, et al. A free-response approach to the measurement and characterization of radiographic-observer performance. J Appl Photographic Eng. 1978; 4:166–171.
79. Egan JP, Greenberg GZ, Schulman AI. Operating characteristics, signal detectability, and the method of free response. J Acoust Soc Am. 1961; 33:993–1007.
80. Chakraborty DP. Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data. Med Phys. 1989; 16:561–568. [PubMed: 2770630]

81. Starr SJ, Metz CE, Lusted LB, et al. Visual detection and localization of radiographic images. Radiology. 1975; 116:533–538. [PubMed: 1153755]
82. Swensson RG. Unified measurement of observer performance in detecting and localizing target objects on images. Med Phys. 1996; 23:1709–1725. [PubMed: 8946368]
83. Obuchowski NA. Nonparametric analysis of clustered ROC curve data. Biometrics. 1997; 53:567–578. [PubMed: 9192452]
84. Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: modeling, analysis and validation. Med Phys. 2004; 31:2313–2330. [PubMed: 15377098]
85. Chakraborty DP. Analysis of location specific observer performance data: validated extensions of the jackknife free-response (JAFROC) method. Acad Radiol. 2006; 13:1187–1193. [PubMed: 16979067]
86. Edwards DC, Kupinski MA, Metz CE, et al. Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model. Med Phys. 2002; 29:2861–2870. [PubMed: 12512721]
87. Obuchowski NA, Lieber ML, Powell KA. Data analysis for detection and localization of multiple abnormalities with application to mammography. Acad Radiol. 2000; 7:516–525. [PubMed: 10902960]
88. Zanca F, Chakraborty DP, Marchal G, et al. Consistency of methods for analysing location-specific data. Radiat Prot Dosimetry. 2010; 139:52–56. [PubMed: 20159917]
89. Chakraborty DP. Validation and statistical power comparison of methods for analyzing free-response observer performance studies. Acad Radiol. 2008; 15:1554–1566. [PubMed: 19000872]
90. Good WF, Abrams GS, Catullo VJ, et al. Digital breast tomosynthesis: a pilot observer study. AJR Am J Roentgenol. 2008; 190:865–869. [PubMed: 18356430]
91. Babinski PJ, Babinski BS. Adaptability through cross-training in radiology departments. Radiol Manage. 2011; 33:45–49.
92. Yablon CM, Wu JS, Slanetz PJ, et al. A report on the current status of grand rounds in radiology residency programs in the United States. Acad Radiol. 2011; 18:1593–1597. [PubMed: 22055800]
93. Gur D, Sumkin JH, Hardesty LA, et al. Recall and detection rates in screening mammography. Cancer. 2004; 100:1590–1594. [PubMed: 15073844]
94. Jiang Y, Metz CE, Nishikawa RM. A receiver operating characteristic partial area index for highly sensitive diagnostic tests. Radiology. 1996; 201:745–750. [PubMed: 8939225]
95. Obuchowski NA, McClish DK. Sample size determination for diagnostic accuracy studies involving binormal ROC Curve indices. Stat Med. 1997; 16:1529–1542. [PubMed: 9249923]
96. Baker SG, Pinsky PF. A proposed design and analysis for comparing digital and analog mammography: special ROC methods for cancer screening. J Am Stat Assoc. 2001; 96:421–428.
97. Zhang DD, Zhou X-H, Freeman DH Jr, et al. A non-parametric method for the comparison of partial areas under ROC curves and its application to large health care data sets. Stat Med. 2002; 21:701–715. [PubMed: 11870811]
98. Wang F, Gatsonis CA. Hierarchical models for ROC curve summary measures: design and analysis of multi-reader, multi-modality studies of medical tests. Stat Med. 2008; 27:243–256. [PubMed: 17340598]
99. Li C-R, Liao C-T, Liu JP. A non-inferiority test for diagnostic accuracy based on the paired partial areas under ROC curves. Stat Med. 2008; 27:1762–1776. [PubMed: 17968858]
100. Pepe, MS. The statistical evaluation of medical tests for classification and prediction. Oxford, United Kingdom: Oxford University Press; 2003.
101. Hadjiiski L, Sahiner B, Helvie MA, et al. Breast masses: computer-aided diagnosis with serial mammograms. Radiology. 2006; 240:343–356. [PubMed: 16801362]
102. Berbaum KS, Dorfman DD, Franken JEA, et al. An empirical comparison of discrete ratings and subjective probability ratings. Acad Radiol. 2002; 9:756–763. [PubMed: 12139089]
103. Jiang Y, Nishikawa RM, Schmidt RA, et al. Improving breast cancer diagnosis with computer-aided diagnosis. Acad Radiol. 1999; 6:22–33. [PubMed: 9891149]

104. Horsch K, Giger ML, Vyborny CJ, et al. Classification of breast lesions with multimodality computer-aided diagnosis: observer study results on an independent clinical data set. Radiology. 2006; 240:357–368. [PubMed: 16864666]

105. Horsch K, Giger ML, Metz CE. Potential effect of different radiologist reporting methods on studies showing benefit of CAD. Acad Radiol. 2008; 15:139–152. [PubMed: 18206613]

106. Armato SG, Roberts RY, Kocherginsky M, et al. Assessment of radiologist performance in the detection of lung nodules: dependence on the definition of “truth”. Acad Radiol. 2009; 16:28–38. [PubMed: 19064209]

107. Miller DP, O’Shaughnessy KF, Wood SA, et al. Gold standards and expert panels: a pulmonary nodule case study with challenges and solutions. SPIE Proc Med Imaging. 2004; 5372:173–184.

108. US Food and Drug Administration. [Accessed November 17, 2011] Radiological Devices Panel. Available at: http://www.fda.gov/ohrms/dockets/ac/08/transcripts/2008-4349t1-05.pdf

109. US Food and Drug Administration. [Accessed November 17, 2011] Radiological Devices Panel: two draft guidance documents for computer-aided-detection (CAD) devices. Available at: http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/MedicalDevices/MedicalDevicesAdvisoryCommittee/RadiologicalDevicesPanel/UCM197419.pdf

110. Obuchowski NA. Predicting readers’ diagnostic accuracy with a new CAD algorithm. Acad Radiol. 2011; 18:1412–1419. [PubMed: 21917487]

111. Chan H-P, Sahiner B, Wagner RF, et al. Classifier design for computer-aided diagnosis: effects of finite sample size on the mean performance of classical and neural network classifiers. Med Phys. 1999; 26:2654–2668. [PubMed: 10619251]

112. Sahiner B, Chan H-P, Petrick N, et al. Feature selection and classifier performance in computer-aided diagnosis: the effect of finite sample size. Med Phys. 2000; 27:1509–1522. [PubMed: 10947254]

113. Kupinski MA, Giger ML. Feature selection with limited datasets. Med Phys. 1999; 26:2176–2182. [PubMed: 10535635]

114. Simon R, Radmacher MD, Dobbin K, et al. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst. 2003; 95:14–18. [PubMed: 12509396]