Evaluation of computer-aided detection and diagnosis systems[a)]

Nicholas Petrick[b)] and Berkman Sahiner[b)]
Center for Devices and Radiological Health, U.S. Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, Maryland 20993

Samuel G. Armato III
Department of Radiology, The University of Chicago, 5841 South Maryland Avenue, MC 2026, Chicago, Illinois 60637

Alberto Bert, Loredana Correale, and Silvia Delsanto
im3D S.p.A. Via Lessolo, 3 - 10153 Torino - Italy

Matthew T. Freedman
Lombardi Comprehensive Cancer Center, Georgetown University, 3900 Reservoir Road, Northwest, Washington, DC 20057

David Fryd
Riverain Medical, 3020 South Tech Boulevard, Miamisburg, Ohio 45342

David Gur
The University of Pittsburgh, Department of Radiology, Radiology Imaging Research, 3362 Fifth Avenue, Pittsburgh, PA 15213

Lubomir Hadjiiski
Department of Radiology, The University of Michigan, 1500 East Medical Center Drive, MIB C476, Ann Arbor, Michigan 48109-5842

Zhimin Huo
Carestream Health Inc., 1049 Ridge Road West, Rochester, New York 14615

Yulei Jiang
Department of Radiology, The University of Chicago, 5841 South Maryland Avenue, MC 2026, Chicago, llinois 60637

Lia Morra
im3D S.p.A. Via Lessolo, 3 - 10153 Torino - Italy

Sophie Paquerault
12300 Village Square Ter., Rockville, Maryland 20852

Vikas Raykar
IBM Research – India G2 Block, 8th Floor Outer Ring Road, Nagawara Bangalore - 560 045, India

Frank Samuelson
Center for Devices and Radiological Health, U.S. Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, Maryland 20993

Ronald M. Summers
National Institutes of Health Clinical Center, Building 10, Room 1C224D, MSC 1182, Bethesda, Maryland 20892

Georgia Tourassi
Oak Ridge National Laboratory, Computational Sciences & Engineering Division, Oak Ridge National Laboratory Oak Ridge, TN 37831

Hiroyuki Yoshida
Department of Radiology, Massachusetts General Hospital and Harvard Medical School, 25 New Chardon Street, Suite 400C, Boston, Massachusetts 02114

Bin Zheng
School of Electrical and Computer Engineering University of Oklahoma 101 David L Boren Blvd, Suite 1001 Norman, OK 73019

Chuan Zhou and Heang-Ping Chan[c)]
Department of Radiology, The University of Michigan, 1500 East Medical Center Drive, MIB C479, Ann Arbor, Michigan 48109-5842

(Received 6 April 2013; revised 6 July 2013; accepted for publication 9 July 2013; published 1 August 2013)

Computer-aided detection and diagnosis (CAD) systems are increasingly being used as an aid by clinicians for detection and interpretation of diseases. Computer-aided detection systems mark regions of an image that may reveal specific abnormalities and are used to alert clinicians to these regions during image interpretation. Computer-aided diagnosis systems provide an assessment of a disease using image-based information alone or in combination with other relevant diagnostic data and are used by clinicians as a decision support in developing their diagnoses. While CAD systems are commercially available, standardized approaches for evaluating and reporting their performance have not yet been fully formalized in the literature or in a standardization effort. This deficiency has led to difficulty in the comparison of CAD devices and in understanding how the reported performance might translate into clinical practice. To address these important issues, the American Association of Physicists in Medicine (AAPM) formed the Computer Aided Detection in Diagnostic Imaging Subcommittee (CADSC), in part, to develop recommendations on approaches for assessing CAD system performance. The purpose of this paper is to convey the opinions of the AAPM CADSC members and to stimulate the development of consensus approaches and "best practices" for evaluating CAD systems. Both the assessment of a standalone CAD system and the evaluation of the impact of CAD on end-users are discussed. It is hoped that awareness of these important evaluation elements and the CADSC recommendations will lead to further development of structured guidelines for CAD performance assessment. Proper assessment of CAD system performance is expected to increase the understanding of a CAD system's effectiveness and limitations, which is expected to stimulate further research and development efforts on CAD technologies, reduce problems due to improper use, and eventually improve the utility and efficacy of CAD in clinical practice. © *2013 American Association of Physicists in Medicine*. [http://dx.doi.org/10.1118/1.4816310]

Key words: computer-aided detection and diagnosis (CAD), computer-aided detection (CADe), computer-aided diagnosis (CADx), performance assessment, standalone performance, reader performance, clinical performance

1. INTRODUCTION

It has long been recognized that clinicians do not always make optimal use of the data acquired by an imaging device.[1,2] The limitations of the human eye-brain system, limitations in training and experience, and factors such as fatigue, distraction, and satisfaction of search may all contribute to suboptimal use of available information.[3–5] Image processing techniques can be applied to medical images in an effort to address some of these issues. Medical image processing attempts to modify the image presented to the readers in such a manner that abnormalities are enhanced for the human visual system. However, image processing alone is unlikely to completely address factors such as fatigue, distraction, or limitations in training. Computer-aided detection and diagnosis (CAD) systems applied to medical images go beyond image processing, such that they may provide specific lesion location information and/or other diagnostic analysis to assist clinicians.

The aim of computer-aided detection (CADe) systems is to mark regions of an image that may reveal specific abnormalities and alert the clinician to these regions during image interpretation. The aim of computer-aided diagnosis (CADx) systems is to provide to the clinician an assessment of disease, disease type, severity, stage, progression, or regression. A CADx system may use image-based information alone or, generally, in combination with other relevant diagnostic data and biomarkers. Some CAD systems may strive to perform both CADe and CADx functions by first identifying potential abnormal regions and then providing a qualitative or quantitative assessment of these identified abnormalities.

Ever-increasing research and development efforts have emerged in the last 25 years to develop and practically implement CAD systems for various types of diseases.[6–13] As a result, a number of CAD systems are commercially available in the United States and worldwide, including, for example, CAD intended for breast cancer detection on mammograms, lung nodule detection on chest radiographs or on thoracic CT images, and polyp detection on CT colonography. The functionalities of CAD systems have also been expanded to various image analysis methods and quantitative tools (e.g., automated segmentation of lesions, size measurement, and dynamic flow information) that may assist clinicians in diagnostic work-up, treatment planning, treatment response monitoring, prognosis prediction, and risk prediction for certain diseases using image-based biomarkers alone or in combination with other biomarkers or clinical information.

To address issues in this important area in medical imaging and diagnosis, the Imaging Physics Committee of the American Association of Physicists in Medicine (AAPM) formed a Computer Aided Detection in Diagnostic Imaging Subcommittee (CADSC) with the charge "to keep the membership apprised of new developments in computer-assisted detection and diagnosis in medical imaging and to develop techniques, practices and standards that address issues in the field as they arise." The CADSC attracts a diverse membership and participants from various sectors (radiologists, CAD industries, academic researchers, and government entities) (see the Appendix). The CADSC formed

subgroups and conducted extensive discussions in four major areas:

- Methodologies for evaluation of stand-alone CAD system performance,
- Methodologies for evaluation of effects of CAD on users—standardization of CAD evaluation technologies,
- Develop QA procedure recommendations for CAD systems implemented in clinical use,
- Develop training and QA procedure recommendations for using CAD systems.

The purpose of this paper is to convey the opinions of the AAPM CADSC members in the first two areas above. The latter two areas are covered in a companion paper.[14] The opinions aim at stimulating further discussions and may serve as a framework for future development of guidelines but should not be interpreted as specific mandates for assessment of CAD systems. The development of CAD systems goes hand-in-hand with their evaluation. CAD assessment is important for a number of reasons, including estimating algorithm performance, establishing its effectiveness for use, and facilitating comparisons among systems with similar intended uses. In addition, proper system evaluation is essential for algorithm optimization during the design and development of CAD systems.

This paper addresses both the assessment of a standalone CAD system (i.e., a CAD system without the end-user) and the evaluation of effects of a CAD system on end-users. Assessment of a standalone CAD system provides information about the expected performance, including confidence bounds, of the computerized system by itself. Evaluation of the effects of a CAD system on end-users is indispensable and critical because, by definition, CAD systems are to be used by a clinician, who is then responsible for making the final decision for each patient.

A number of study factors can affect the accuracy and precision of performance assessment. These include:

- the selection of the training and test data sets for system design and evaluation;
- the method for determining the reference standard, i.e., deciding which cases in the data set contain the target disease, and identifying the location and extent of disease when present;
- the mark-labeling criterion used for deciding which CADe (or reader) marks correctly point to the diseased locations (true positives or TPs) and which point to nondiseased locations (false-positives or FPs);
- methodology and metrics for assessing standalone CAD system performance;
- the design of observer performance experiments; and
- methodology and metrics for assessing the clinical impact of a CAD system on clinicians.

The first three factors listed above can affect the performance assessment in different and important ways. For each of these factors, the implementation and the effects are likely similar for the two types of assessment considered in this paper, i.e., the performance assessment of standalone CAD systems or that of clinicians aided by the system. For example, one needs to determine which cases in the data set contain the target disease regardless of the assessment type, and the method used for establishing the presence of disease can have a similar effect on both types of performance assessment. Therefore, the discussion below pertaining to these three factors applies to assessment of both standalone CAD and readers with CAD. The evaluation methodologies for CADe and CADx systems are similar in many aspects, but they also differ in some critical ways. These important differences will be noted. In addition, although it is impossible for this paper to cover the specific evaluation methods for each type of CAD applications, the principles and some basic approaches addressed herein should be common to and may provide some guidance for the development of CAD systems in general.

2. DATA SETS

Development and performance assessment of CAD systems typically require the use of patient images with and without the target disease. The collections of patient images, together with other relevant diagnostic data and biomarkers, used for CAD development and subsequent performance assessment are typically referred to as training and test data sets, respectively. The target population is defined as the population of patient cases to which the CAD system is expected to be applied. A properly selected training data set allows the developer to choose and fine-tune the structure and parameters of a CAD system to the expected properties of disease-positive and disease-negative cases in the target population. A properly selected test data set allows for estimation of CAD system performance and facilitates performance comparisons among CAD systems that have similar intended uses.

2.A. Essential patient data

In the field of medical imaging, the main component of a data set for a CAD system is human images, although other supporting data are also essential. For example, for CAD devices targeted for radiology applications, results from another diagnostic test such as pathologic evaluation are critical for determining the reference standard. For a CADx system to achieve its full clinical potential, incorporation of nonimaging biomarkers and other relevant patient-specific information, such as patient age, demographics, family history, concomitant diseases, and environmental exposure history, may be necessary. In addition to patient-specific information, lesion-specific information such as disease type and lesion size is often needed for defining the set of patients for primary and/or sub-group analyses. The image data and any available relevant information for a given patient can be collectively considered as a case sample in a data set. Although images are often used as example in the following discussion for data collection, the requirements such as case composition, sample size, data verification, and measurement standardization

are generally applicable to other types of data in the data set.

2.B. Data set types and composition

CAD systems invariably contain a set of parameters, such as those for image enhancement, thresholds for selecting object candidates, variables used in segmentation, features selected for false-positive reduction, and classifier weights. Although it is possible at times to select some of these parameters based on the general knowledge of disease appearance, a set of cases is usually required for selecting and/or optimizing such parameters. The data set used for system design is called the training data set. The performance estimate obtained by applying the CAD system to the training data set is termed the resubstitution estimate, which is usually optimistically biased.[15] Ideally, an independent data set is used for estimating the performance of the CAD system or that of readers using the system. The data set used for system assessment is called the test data set. In practice, there are also data sets that lie somewhere between training and test with regard to their use in the design and intermediate assessment of a CAD system. For example, a data set may be used for evaluation within the system development phase, with the results used as a guide for system optimization. Following the nomenclature used in the fields of statistical machine learning, pattern recognition, and neural networks, we refer to the data sets used in these intermediate assessments as "validation" data sets.[16–18]

2.B.1. Training data set

To properly train a CAD system, the training set should include the range of verified abnormalities or features of interest as seen in practice (e.g., different sizes, shapes, locations, and levels of conspicuity) for the expected range of patient characteristics (e.g., age and comorbidity). This range should be consistent with the target population for the CAD device. A variety of cases without the abnormality of interest should also be included, if such cases are available and if the inclusion of such cases is appropriate for the task under investigation. The presence of other concomitant diseases or abnormalities should be detailed.

Ideally, the training set should be large enough to allow the estimation of the required parameters, and representative of the underlying probability distributions in the target population. In practice, however, due to sometimes unavoidable biases in the data collection process, some patient subgroups may be under- or over-represented. At other times, over-representation of certain patient sub-groups may be deliberate if it is believed that over-representation of such patients is beneficial to CAD development, perhaps because the detection or classification task is particularly challenging for the sub-group. The effect of the representativeness (or lack thereof) of the training set on the resulting performance of the CAD systems is a research topic that requires further investigation. Because the composition of the training set can have a significant effect on the performance of the CAD system, authors of publications should clearly describe relevant characteristics (e.g., image acquisition parameters, lesion size and shape, etc.) of their training sets, and paper reviewers should consistently demand these details if they are not provided.

A training set may also include: (1) cases that simulate the target disease (e.g., benign tumors, mimickers of malignancies, etc.); (2) images of models or phantoms; (3) simulated lesions or disease patterns superimposed on disease-free patient images; (4) electronic addition or removal of disease or change of location;[19,20] or (5) modification of disease appearance produced by image processing so that the abnormality is altered (e.g., in size, shape, rotation, edge definition, density, location, etc.). Images or abnormalities that have been obtained or modified by image processing should have undergone rigorous evaluation to verify that they have characteristics similar to actual abnormalities depicted on patient images. A CAD system trained with simulated images that do not appear authentic may not be generalizable to real patient images.

2.B.2. Test data set

Ideally, the composition of the test data set should match the target population to which the CAD system is intended to be applied. Also, image acquisition and patient preparation parameters should ideally be representative of those found in the target population. To allow proper interpretation of test results, inclusion and exclusion criteria must be clearly stated and must be justified as necessary. The distribution of the known covariates (e.g., lesion type and size, disease stage, organ characteristics, patient age, etc.) should be specified, and any significant departure from those of the target population should be identified and discussed in the publications that report the result of performance assessment.

For the purpose of performance assessment, especially when disease prevalence is low, CAD data sets are often enriched to include a larger proportion of diseased cases compared with the clinical disease prevalence in the target population. To enrich the prevalence, one option is to collect a consecutive set of diseased cases and a consecutive or randomly selected set of nondiseased cases from the target population. Mathematically, this type of prevalence enhancement does not affect the estimates of sensitivity, specificity, and the area under the receiver operating characteristic (ROC) curve, but does impact performance metrics such as positive and negative predictive values. In addition, it has been shown that prevalence enhancement can affect reader behavior in observer performance studies.[21–23] Sometimes, when two competing modalities are compared (e.g., two competing CAD systems, or interpreting cases with or without the aid of CAD), the test data set may include a larger proportion of cases that accentuate the difference between the modalities. This type of enhancement, referred to as stress testing, is covered in more detail below (Sec. 2.C).

Test data set selection can be a major source of bias in CAD system assessment, which is a risk shared by many

diagnostic tests. Selection or spectrum biases are introduced when selected cases for a study are not representative of the target population.[24] Verification bias, among other additional sources of bias, occurs when a study is restricted to patients with definitive verification of disease status.[25] For certain diseases and modalities under investigation, it may be possible to reduce these biases by carefully planning the study and case selection, or by selecting consecutive cases from the target population, but some amount of bias from case selection will be unavoidable in many studies. It is important to acknowledge and discuss potential biases in the data set composition to allow proper interpretation of study results.

2.C. Special types of test data sets

2.C.1. Stress testing

The purpose of a stress test is to study differences between competing systems or reading modalities using cases selected specifically to challenge those differences.[26] Because of the special selection, these cases are not expected to be present in a similar proportion as that in the target population. As a result, performance estimates for each modality may be biased, and are therefore unlikely to exactly match the performances in the target population. In the context of CAD, the goal of stress testing is not to evaluate CAD performance on an absolute scale, but to compare the relative performance of two CAD systems, or clinician performance without and with CAD in a specific scenario. For example, if it is expected that the difference between radiologists interpreting images without and with CAD is in cases containing small noncalcified lesions, a test data set enriched with cases containing small noncalcified lesions may be appropriate.

Not all types of enrichment is stress testing, and not all types of enrichment are appropriate for CAD assessment. For example, when evaluating the effect of CAD on radiologists, it is not appropriate to include only the subset of cases where the CAD system has a smaller than average number of FPs. Even if one observes an improvement with CAD for this subset of cases, the results may not be generalizable to the target population of cases because radiologists' performance may deteriorate with CAD for cases that have a larger than average number of FPs.[27]

2.C.2. Test data set reuse

Sample size is a major factor that affects the variance of performance estimates, and, given the difficulties of collecting a large and representative data set, it is tempting to reuse the same test data set multiple times for CAD assessment. However, this practice creates the risk of tuning the CAD algorithm implicitly, or even explicitly, to the test data. For example, with the knowledge that the CAD system failed to provide correct result for a given test case, it may be relatively easy to find a remedy that may solve the problem for this specific case. Even if the CAD developer is not informed of the results of individual cases but is allowed to test the CAD system on a given data set unlimited times and informed of the collective performance every time, the CAD system could still be tuned to obtain high performance for the test set eventually. However, such performance may or may not be generalizable to the target population. In these situations, the integrity of the test data set as an independent set may be compromised. Using a completely new data set for each evaluation study reduces or eliminates this bias, but it will cause a substantial cost increase in CAD development and assessment. In addition, it will be difficult to increase the size of the test data set to reduce the variance in the performance estimates if old cases have to be removed.

Research is needed to develop techniques that balance the variance reduction in reported performance estimates with the potential bias associated with data reuse. One approach might be to add newly collected cases to an older test data set allowing the test data set size to grow (i.e., reduce the variance) and then utilize a random sample from this larger data set. This approach still has the potential to bias the results due to repeated use of some of the test cases. Restricting the developers of CAD algorithms from having direct access to the data and detailed knowledge of performance may reduce, but not eliminate, this bias. In general, when data reuse is allowed in the process of CAD assessment, techniques should be incorporated to ensure the integrity of the test data by demonstrating that the reused data does not introduce unreasonable bias.

2.C.3. Common sequestered test data sets

Patient image data sets and relevant auxiliary data (e.g., reference standard, lesion location and extent, imaging equipment and acquisition parameters, and patient demographics), as well as other relevant biomarkers if available, accessible to the scientific community are essential for increasing the pace of CAD research and development. Such data sets may help researchers navigate an important barrier to the initiation of a new CAD project, i.e., the collection of a large, representative and adequately verified database. For established CAD applications, a properly administered common database may facilitate the performance comparison among CAD systems or a comparison of previous and current versions of the same CAD system.

Currently, only a small number of such data sets are available for CAD development.[28–30] CAD developers can use these data for training, validation, and testing purposes. Therefore, evaluation results reported on public data sets may be anywhere between an estimate of resubstitution performance to true test performance. A possible solution to this problem may be to have an independent laboratory conduct the testing by using a sequestered data set. Clearly, the issues discussed above related to reusing test data will need to be addressed so that the CAD systems will not be trained to the sequestered test set and lead to biased performance estimates. Measures to reduce bias should be implemented such as using only a random subset of the sequestered data set for each

test of a given CAD system, providing the CAD developer only the overall results (e.g., sensitivity, specificity, or possibly a performance curve), and continuing to collect new cases to the sequestered test set. More research is needed to study the potential bias on performance estimates with the reuse of a sequestered data set, and how to minimize this potential bias.

We believe that establishing an independent laboratory to oversee the collection of sequestered data sets, and to conduct independent testing, as well as further development of public data sets for CAD system training for various clinical tasks will accelerate the advancement of CAD.

3. REFERENCE STANDARD

CAD device development and evaluation generally rely on databases of medical images with known disease status for each patient, and marked location and extent of the disease within each image. We use the term "reference standard" to indicate the established factual information regarding patient's status for the target disease and other relevant information such as the location and extent of the disease. For clarity, a case-level and a lesion-level reference standard should be defined separately. The case-level reference standard is the determination of the disease status for each patient. The case-level reference standard is often all that is required for conducting certain types of clinical CAD evaluations. The lesion-level reference standard includes a determination of disease status as well as the location and extent of the disease for all individual lesions, which are important, for example, for evaluating the capability of a CADe system in alerting the clinician to specific lesion locations or a CADx system in quantifying lesion sizes for characterization or monitoring treatment response.

3.A. Performance assessment with a "gold" reference standard

The disease status of a patient or lesion is often obtained through an established clinical method, usually relying on a determination from an independent device or procedure; through follow-up clinical examinations; or through the interpretation of the image data by reviewing clinician(s). Ideally, the reference standard is a perfect gold standard, which is also sometimes referred to as the ground truth. Although the formation of a perfect gold standard is rarely if ever achievable in practice, the performance assessment of many CAD devices is often based on such an assumption. An example is the use of biopsy and pathologic examinations as the reference standard for the presence of cancer in radiological images of a patient or region. Biopsy is often considered the "gold standard", even though the reported frequency of errors for anatomic pathology could range from 1% to 43% of specimens.[31] The reference standard for positive cases is generally based on the pathology report for the presence of disease. The reference standard for negative cases is generally established by biopsy or follow-up over an appropriate time interval to demonstrate that the case is negative for the target disease. Examples of other gold reference standards include the use of conventional catheter angiography for coronary plaques and optical colonoscopy for polyps in CT colonography.

When the location and extent of the disease are required (e.g., for standalone evaluation of a CADe system), the task reverts to having experts (i.e., experienced clinicians specifically tasked with defining parts of the reference standard) review the imaging study to identify the location and extent of the lesions by using any available clinical and pathology reports and follow-up information for verification. This process can be quite challenging and is in many aspects similar to the lack of a gold standard scenario discussed below. It may be necessary to match a lesion on one reference modality with the same lesion on the images to which CAD will be applied (e.g., matching lesions found in reference optical colonoscopy with their presentation in CT colonography). It is also important to verify that the disease is visibly depicted in the modality to which CAD will be applied, because there is generally no interest in highlighting totally occult abnormalities that can only be detected by a different procedure or modality. Often, a single clinician can perform this task, but inter- and intrareader variability exist in defining the location and extent of a lesion. This variability is likely to depend on the imaging modality, disease, and availability of other sources of information in addition to the clinical images. Although it is difficult to provide a general recommendation on this complex issue, one acceptable, although still imperfect, solution may be to have multiple clinicians perform this task and develop a combined reference standard for lesion location and extent (see Sec. 4).

Clinicians participating in the truthing process should not participate as users for performance assessment of the CAD device because doing so might introduce bias into the study results.

3.B. Performance assessment without a "gold" reference standard

It may sometimes be difficult or impossible to establish a gold reference standard. As an example, CT pulmonary angiography has become the standard of proof for pulmonary embolism (PE) in a patient, but there is no independent test for verifying the location of most PE seen on thoracic CT images. Therefore, while clinical evidence may be used to establish a gold standard for the disease status, it is not yet possible to establish a reliable gold standard for specific PE locations to be used in CADe for PE on CT pulmonary angiography. As a second example, while pathology may serve as the reference standard for other CAD systems, pathology CAD systems (i.e., CAD algorithms designed to aid in detecting or diagnosing disease on pathology slides) may not have a readily available independent test to serve as the reference standard.

When a gold standard is lacking, one common approach is to establish a reference standard by having an expert panel

of clinicians interpret the images along with other available clinical information (e.g., the radiology report).[32–36] and combining these expert interpretations. One approach for combining expert interpretations is to have each expert first independently review the information and make an initial determination on his or her own. After their initial readings, an adjudication method (e.g., majority opinion, independent arbiter, or a statistical technique such as latent class analysis[36]) may be utilized for defining a binary reference standard.

As a general rule, when a new technology has a higher sensitivity than an existing technology, abnormalities detected by the new technology may be scored erroneously as FPs if the reference standard is based only on the existing technology. Considering CAD as the new technology, review of CAD marks by the expert panel members may therefore potentially be helpful in determining the reference standard in the absence of a gold standard. However, this potential should be weighed against the risk of inflating the performance of the specific CAD system used to assist in determining the reference standard, and the risk of underestimating the performance of other CAD systems that may be evaluated with the resulting reference standard. In any case, the panel should not rely only on CADe marks in the truthing process.

An important consideration in conducting subsequent evaluations that rely on imperfect truth (e.g., panel-based truth) is that the uncertainty in the truth leads to additional uncertainty in the observed results.[37] This additional uncertainty should be accounted for in the final analysis whenever possible.[32]

In general it is very difficult to achieve a perfect reference standard in many situations, as even pathology is imperfect. On the other hand, the needed level of "truthing" will depend on the objective of the study, and imperfect but perhaps sufficient information may be used to establish an adequate reference standard for the intended purpose of assessing CAD performance. For a given data set, the method of establishing the reference standard should be clearly described.

4. MARK-LABELING

The performance of a CADe system often depends not only on whether it can correctly identify a diseased case but also whether it can correctly locate the abnormality in the patient images. Mark-labeling is defined as the set of rules used for deciding which marks correspond to the targeted abnormalities. These rules establish the amount of correspondence required between the CADe (or reader) marks and the disease location/extent in the reference standard. A mark is considered a TP if it satisfies the mark-labeling rule, and an FP otherwise.

A wide range of mark-labeling rules has been used by CAD researchers. The following, either alone or in combination with others, are some of the rules used most often for defining TPs:

(1) There is an overlap in area between the mark and the reference standard annotation,[38,39] and this overlap divided by the union of the two areas exceeds a preselected overlap threshold,[40] or this overlap divided by the area of the reference standard exceeds a preselected overlap threshold.[41–43]
(2) The distance between the centroid of the CADe (or reader) mark and the annotated reference standard is less than a preselected distance threshold,[44–47] or a distance threshold that depends on the size of the lesion.[48,49]
(3) The center of the mark is within the annotated abnormality,[50–52] or vice versa.[53]
(4) Visual evaluation by an expert or a researcher indicates that the mark points to the annotated abnormality.[51,52,54]

The choice of the mark-labeling rule can have a major effect on the reported performance of the CADe system. Kallergi *et al.*[55] studied different rules for mark-labeling, including some of the methods mentioned above, and found that different criteria can result in dramatically different TP and FP estimates. It is therefore critical that the mark-labeling rules used for standalone and clinical assessment for a CADe system be described, and rationale for choosing such a method be provided.

Mark-labeling by a human expert might be a more expensive approach compared with objective rule-based mark-labeling. In addition, mark-labeling by a human expert does not lend itself to an objective comparison of different techniques developed by different groups because it is likely to be more subjective and will involve intra- and interobserver variations. However, since the purpose of CADe marks is to improve the detection and perception of abnormalities by human readers, a human reader may be a good judge for deciding whether a mark points to the true lesion. If this approach is used, the mark-labeler should not be a participant in the clinical assessment of the CADe system, should not have a stake in the outcome of the assessment, and should be provided with both the mark and the reference standard annotation. Moreover, the expertise level of the mark-labeler, any instructions to the mark-labeler, and specific criteria used as part of the mark-labeling process should be documented along with the reported results. When multiple mark-labelers are involved, the process by which their interpretations are combined to make an overall mark-labeling decision should be described.

An objective rule-based approach for mark-labeling will generally be consistent; however, the mark-labeling rule should be carefully selected so that computer/reader marks are consistent with clinical interpretations. Marks labeled as TPs that are unlikely to attract the attention of the reader to the abnormality are particularly worrisome because they may lead to inflated standalone CAD performance estimates. For example, a large mark that covers a large portion of the breast has a higher chance to overlap with a reference mark but the CAD system may actually have missed the true lesion. Extreme caution is therefore warranted for rules that allow large marks to be used for TP labeling. The use of these types of rules should be accompanied by additional statistics

and analyses to demonstrate that the rules are reasonable. In general, the areas of the TP marks are expected to be comparable to the lesion sizes as determined by the reference standard.

A large fraction of CADe publications do not describe the mark-labeling rules used in system assessment. We selected a random sample of 58 CADe publications on nodule detection in thoracic CT published in the last ten years; 47 of them did not describe the mark-labeling protocol. In a similarly selected sample of 21 publications on polyp detection in CT colonography, 9 did not describe the mark-labeling method. It is impossible to put into proper perspective the assessment results of a CADe system if details of mark-labeling are missing or incomplete. Authors of publications should clearly describe the mark-labeling method used in CADe studies, and reviewers should consistently demand details of the mark-labeling process if they are not provided.

5. CAD ASSESSMENT METRICS

This section discusses the methodologies and metrics that are available for measuring CAD performance, either standalone or when the CAD system is used by a clinician. Most of this section focuses on CAD assessment for a binary reference standard which contains only two possible truth states (e.g., diseased/nondiseased or lesion/nonlesion). The cases that contain a lesion will be referred to as *actually positive cases*, and those that do not contain a lesion will be referred to as *actually negative cases*. The multiclass problems (i.e., more than two possible truth states) are briefly discussed in Sec. 6.D.

5.A. Basic definitions

When both the reference standard and the diagnostic test results are binary, the sensitivity (or TP fraction, TPF) of the test is defined as the percent correct on the set of actually positive cases, and the specificity is defined as the percent correct on the set of actually negative cases. The FP fraction (FPF) is defined as 1-specificity. TPF and FPF are properties inherent to the diagnostic test, and are independent of disease prevalence, so that the results from a prevalence-enriched test data set are representative of the true TPF and FPF of the test.

The positive predictive value (PPV) is defined as the proportion of positive calls by the test that are TPs, and the negative predictive value (NPV) is the proportion of negative calls that are true negatives (TNs). These two measures can be expressed in terms of TPF, FPF, and disease prevalence. Because prevalence is incorporated, PPV and NPV are not properties of the test alone, but the results of applying the test to a certain data set or target population.[56]

5.B. Metrics based on ROC analysis

Most CADx systems provide a multilevel (i.e., more than a binary or two-level) rating for clinicians to consider as part of their interpretation. For such systems, and when the reference standard is binary, ROC analysis is a common choice for standalone performance assessment. Likewise, ROC analysis is typically appropriate when the task of the clinician is to provide an ordinal rating for the likelihood of disease in a case. The ROC methodology analyzes ratings by labeling them according to the truth state and the comparison of the rating to a decision threshold. To obtain an empirical ROC curve, the threshold is varied to cover the entire range of possible ratings, and the TPF is plotted as a function of the FPF. The specific (TPF, FPF) pair at a given decision threshold is termed an operating point.

An advantage of the empirical ROC curve is that no structural assumptions are made about the form of the plot and the underlying distributions. However, an empirical ROC curve is not smooth, particularly when the data and/or the number of levels in the rating scale are limited. Parametric ROC methods model the data for actually positive and actually negative cases so that a smooth ROC curve is obtained. However, any incorrect assumption related to the underlying model would impact the quality of the fitted ROC curve. The binormal model[57] was one of the earliest parametric approaches for fitting an ROC curve to raw data. Later approaches, such as the bi-gamma,[58] proper binormal,[59] and contaminated[60] ROC models, were mainly aimed at addressing some of the undesirable artifacts (e.g., nonconvexities or degeneracies) caused by the earlier models.

Various figures of merits (FOMs) can be estimated based on the fitted ROC curves, including the AUC, the partial AUC[61] (area under just a portion of the ROC curve), and a sensitivity/specificity pair at an operating point. These FOMs can also be estimated directly from the raw ROC curve by nonparametric methods.[62] Both parametric and nonparametric methods have been developed for estimating the uncertainties in ROC curves and FOMs. These methods have been further extended for statistical comparison of two (or more) systems or readers. An extensive body of publications exists for the development and application of parametric and nonparametric approaches to ROC analysis. We can refer only to a limited subset here.[26,63–72] Articles by Wagner *et al.*[26] and the International Commission on Radiation Units and Measurements (ICRU) (Ref. 71) provide a comprehensive summary and explanation of these assessment methodologies and also refer to other prior publications.

When the ratings are multilevel and the reference standard is binary, ROC methodology has advantages over a sensitivity/specificity pair determined at a single decision threshold. For example, in many studies comparing clinicians' performance with and without CAD, the sensitivity at a selected threshold may be higher when the clinician is aided by CAD, but the specificity may be higher without CAD. When the ROC curves do not cross, the AUC difference, which represents the average sensitivity over the entire specificity range, may be used to rank the two modalities, with confidence intervals that depend on the magnitude of the difference and the variability of the AUC estimates. The use of the AUC not only eliminates the need to use an arbitrary threshold, but may also reduce the uncertainty of the FOM compared

to a sensitivity/specificity pair. The AUC, partial AUC, and a clinically relevant sensitivity/specificity pair can all be reported as metrics depending on the purpose of the study. However, it may be appropriate to report metrics relevant to binary ratings (e.g., sensitivity/specificity pair, PPV/NPV pair) in prospective studies that evaluate a CAD system in clinical settings.

Finally, it is important to predefine which FOM will be used as the primary endpoint of a particular study, because different FOMs may or may not lead to the same sample size estimate (through power analysis) or study conclusion.

Dedicated software has been developed for estimating ROC-based FOMs.[73–75] When using the software packages or any other methods to estimate FOMs, uncertainties (e.g., standard deviations or confidence intervals) should always accompany the point estimates. It is a good practice to inspect the ROC curves even when the primary purpose of the ROC analysis is to report an FOM, because a visual inspection can easily discover artifacts or crossing curves when two or more ROC curves are compared.[76]

5.C. Metrics based on location-specific ROC analysis

The data required for location-specific ROC analysis consist of the identified locations of suspected abnormalities and a rating for each abnormality. These data are generally available in standalone CADe assessment and can be collected in many reader performance studies involving CADe. Each rating-location pair is labeled as a TP or FP depending on the location of the computer/reader mark, the rating, and the truth state.

In the localization ROC (LROC) paradigm, the most suspicious location in each image is marked and rated. The LROC curve plots the fraction of actually positive images with a correctly localized lesion as a function of the FPF for actually negative images.[77] In the free-response ROC (FROC) paradigm, the number of mark-rating pairs for an image is not constrained. The FROC curve plots the fraction of correctly localized lesions as a function of the average number of FPs per image.[78]

For LROC data, a model and fitting procedure have been described.[79] In order to fit FROC data, some researchers relied on parametric ROC methods,[80] while others developed a visual search model[81] and used a maximum likelihood procedure.[82] Chakraborty also developed a jackknife FROC (JAFROC) method, available as dedicated software,[83] that provides an FOM to summarize FROC data and statistically compares the FOMs of competing systems.[84,85] Other approaches for defining and analyzing location-specific FOMs include: an approach that penalizes for the number of FP marks, rewards for the fraction of detected abnormalities, and adjusts for the effect of the target size;[86] the use of bootstrapping to estimate the uncertainties in the FROC FOMs;[87,88] and the use of an exponential transformation for the number of FPs per image (EFROC curve) to overcome the theoretically infinite limit to the area under the FROC curve.[89]

Simulation studies have been conducted to compare some of the location-specific ROC methods and to compare them to ROC analysis.[90] The results demonstrated that FROC is statistically more powerful than ROC.[84,90] From a statistics point of view, location-specific ROC methods may be more appropriate when location data are available. However, in studies involving observers, the data collection and analysis methods (with or without localization) will depend on the specific task and the relevance of the task to clinical practice. In situations where data collection without location is clinically more relevant, ROC methods may still be more appropriate.

6. STANDALONE CAD ASSESSMENT

Standalone performance is an important measure for a CAD system as it may reveal the magnitude of the potential impact (or limitation) of the system at an early stage prior to testing of reader performance using the CAD system. It also provides an assessment of the system performance independent of the human factors and can be used to efficiently study the system performance on subgroups of the population. Standalone performance assessment studies can often be conducted on larger data sets compared to reader studies, which are considerably more burdensome to perform on large data sets. Standalone performance assessments can therefore provide a more detailed description of the algorithm performance compared with reader studies.

6.A. ROC or location-specific ROC performance assessment

A CAD system may provide either a multilevel rating or a binary output for the user. When a multilevel computer output is available, the standalone CAD performance can be analyzed using ROC or location-specific ROC methods. Many CAD systems that provide only a final binary output actually convert a multilevel output to a binary value by selecting an operating point at the end of system development. For these systems, it is logical to assess the standalone CAD system based on its performance over the entire range of potential thresholds, before the operating point is selected, by using ROC or location-specific ROC analysis. For the assessment of standalone CADe systems, location-specific ROC is more appropriate than ROC since these systems provide marks corresponding to computer-detected locations of lesion candidates. As discussed below, it is also critical to assess such systems at the selected operating point.

6.B. Performance assessment for CADe at the selected operating point

The output of a CADe system is usually displayed to the end-users as prompts at a given decision threshold or device operating point. Therefore, another relevant metric for an end-user is the performance of the CADe system at the chosen operating point. At a minimum, the sensitivity and the mean

number of FPs per case should be reported with accompanying confidence intervals. The performance at an operating point can be further stratified by lesion sub-type, size, case type, or image acquisition parameters for a more in-depth understanding of standalone performance.

When multiple, independently acquired views of an organ are available for CADe (e.g., in mammography and CT colonography), or when the images contain multiple abnormalities, different definitions of sensitivity are possible. For example, one can define a "lesion-based" sensitivity for which each missed lesion constitutes a false-negative (FN), a "view-based" sensitivity for which only actually-positive views with no correctly localized lesions constitute FNs, or a "case-based" sensitivity for which only actually-positive cases with no correctly localized lesions constitute FNs. To avoid confusion, it is important to clearly define the type of sensitivity used in a standalone CAD assessment study.

A commonly used method for estimating the confidence interval for a fraction such as the sensitivity is based on the normal approximation (also known as the Wald interval).[91] However, for small sample sizes and near the endpoints, the normal approximation performs poorly and is not recommended. For computing the confidence interval, either the Jeffreys interval[92,93] or the Agresti-Coull[94] interval is recommended.

The mean number of FPs (per case or per view) is generally assumed to be a random variable with a Poisson distribution[95,96] and is estimated using the sample mean ($\hat{\lambda}$). Chakraborty and Winter[97] derived the confidence intervals by using the standard normal approximation and by using the fact that λ is a scaled version of the FPF. If the normal approximation is inadequate, the Jeffreys interval[93] is recommended.

The confidence intervals recommended above assume that the observations (i.e., CAD system outputs) are independent. Appropriate modifications have to be made if they are not independent. In most CADe applications, the CADe marks will be dependent because there may be multiple lesions or marks in the same case, and CADe ratings for multiple lesions or regions in the same case are typically correlated. Such data are referred to as clustered data.[98] If the correlations are positive, the variance will be underestimated by the conventional binomial method, which assumes independence. To correct for this underestimated variance, a ratio estimator for the variance of clustered binary data has been derived.[99] Alternatively, bootstrap methods can be used for computing the confidence intervals.[100]

6.C. Nature of FPs and FNs

Characteristics of FPs and FNs are important additional information related to the standalone performance of a CAD system.[101] Analysis of these characteristics for a particular CAD system will help developers target these types of errors in their future work. In addition, the characteristics might provide information on whether some FPs are easy for a clinician to dismiss and whether some FNs are easy for the clinician to detect or diagnose even without CAD.[102] As an example, a good CADe system designed for a second-reader mode application should point out inconspicuous lesions that, when highlighted, can be identified as disease by a clinician. It may be acceptable if a CADe system misses lesions that are easy to detect because a clinician will be likely to detect them in his/her unassisted interpretation.

6.D. Multiclass classification

The assessment methods described above are designed for classification tasks for which the reference standard is binary. For some tasks, the reference standard may contain more than two truth states (or classes), and the CAD system may be designed to distinguish among more than two classes. A number of methods have been proposed in the literature for the evaluation of this multiclass problem. One approach is to break the problem into multiple two-class classification tasks, and to use ROC methodology to analyze these multiple two-class problems.[103,104] Another approach is to treat the problem as multiclass, but consider only a fixed operating point. Yet another approach is to generalize ROC analysis to full multiclass problems.[105–110] Although FOMs for multiclass ROC analysis have been defined under certain restrictive assumptions, a general FOM remains a subject of ongoing research.

7. ASSESSMENT OF READER PERFORMANCE WITH CAD

CAD devices are specifically designed to aid the clinician in the decision making process, so the clinician is an integral part of the evaluation of their effectiveness. For simplicity, we use the term "reader performance assessment" to denote a test (study) designed to evaluate the performance of a clinician using CAD as part of the decision making process, regardless of whether the CAD system is an aid for image interpretation or other diagnostic purposes. Reader studies are more indicative of the clinical effectiveness of a CAD device compared with standalone testing.

If a single reader participates in a reader study, many of the methods described in Secs. 6.A and 6.B are applicable. For example, if the reader provides a multilevel rating, the reader data can be analyzed using ROC or location-specific ROC methods. If the reader provides a binary rating, then sensitivity, specificity, and their confidence intervals can be estimated using methods described in Sec. 6.B. However, as described below, single-reader studies are generally not sufficient for CAD assessment as they do not capture reader variability, and their results are typically not generalizable to the population of readers targeted for the device. We discuss the generalizability of CAD reader performance studies and analysis techniques next.

7.A. Generalizability of reader performance studies

An important consideration in the design of any performance study is the generalizability of the results. Generalizability is a design concept that allows a system to be evaluated

in such a way that the results from the study can be generalized to a wider population. Ideally, the results of a reader performance study should generalize to the targeted populations of patients examined, the clinicians interpreting the images, the imaging hardware utilized, and the reading environment. However, developing a globally generalizable study under controlled conditions can be cost-prohibitive. Instead, a controlled study is limited to sampling influential groups or factors such that the study results can be generalized only to these specific, but limited populations or conditions.

In general, two important factors in CAD assessment are the population of patients undergoing the examination (cases) and the population of clinicians interpreting the image data.[26] Both of these have been found to be major sources of variability in medical image interpretation. While it is generally recognized that differences in cases add variability, the potentially large variability from a population of interpreting clinicians is sometimes overlooked. A classic demonstration of reader variability is found in the work of Beam *et al.*[111] This study included 108 mammographers in the United Stated reading a common set of mammograms and showed that readers had a 40% range in sensitivity, a 45% range in specificity and an 11% range in radiologists' ability to detect cancer as indicated by the AUC.[111] The study showed both a range in reader expertise (i.e., some mammographers performed better than others in terms of a higher ROC curve) and a range in reader operating points (i.e., the mammographers self-calibrated to different tradeoffs between sensitivity and specificity in terms of operating at different points along an ROC curve). Both of these factors are important and may contribute to overall reader variability.

Multireader multicase (MRMC) study designs have been developed to appropriately account for both reader and case variability, and they are the most practical designs for evaluating reader performance with CAD. Other factors (e.g., patient preparation, acquisition parameters) may need to be accounted for in specific types of CAD devices, but at a minimum both reader and case variability should be addressed when CAD is assessed.

7.B. Multireader multicase (MRMC) analysis methods

In MRMC reader studies, a set of readers interprets a common set of patient images under all competing reading conditions (e.g., unaided readers versus readers aided by CAD). This study design is preferred for evaluating the impact of a CAD on reader performance because it accounts for both reader and case variability and has been used by a number of researchers in their CAD studies.[27,112–115] For MRMC studies of CADe, the style of the mark used to identify an abnormality (e.g., arrow point to or circle surrounding the abnormality) may impact reader performance.[116] Consistency in the mark style across CADe algorithms in a comparison study is recommended to minimize the impact of any confounding effect associated with a difference in mark styles.

An MRMC study is called "fully crossed" when all participating readers independently read all of the cases with all modalities. This design offers the most statistical power for a given number of truth-verified cases.[26] While the fully crossed design is the most common approach in retrospective reader studies, other hybrid MRMC study designs have been proposed[117,118] which may be more efficient in time and cost compared to a fully crossed MRMC study.

Various groups have developed approaches to analyzing the statistical significance of effects and for sizing MRMC studies based on pilot data. These methods vary in their assumptions.[119] Some of the most common MRMC methods include the Dorfman-Berbaum-Metz (DBM) method,[120] the Obuchowski-Rockette (OR) method,[121,122] and the one-shot MRMC variance estimate.[123]

The DBM method is based on an ANOVA analysis of jackknife pseudo-values,[120] while the OR method employs ANOVA, but with a modified F-test to correct for the correlations between and within readers.[121] The one-shot MRMC method is based on the mechanistic MRMC variance and the nonparametric variance of a single-reader U statistics AUC estimate.[123] All of the methods handle a range of performance metrics (e.g., AUC, sensitivity, specificity), accommodate nonparametric estimates, and have software available for download.[73–75,86,124] The DBM and OR methods also accommodate parametric models. In addition, Chakraborty and Berbaum developed a JAFROC method to analyze an alternative free-response ROC figure of merit in MRMC studies as discussed above.[84]

7.C. MRMC study designs

7.C.1. Prospective CAD studies

Reader performance testing can be conducted through a prospective reader study or field test (e.g., a randomized controlled trial) evaluating the CAD system under actual clinical conditions. Prospective studies of CAD are uncommon because they require more coordination to ensure that patient care is not adversely affected, typically take longer for accrual of patients, and generally require a much larger patient population, especially when the prevalence of the target disease is low (e.g., breast cancer screening).

Prospective evaluations of CAD typically fall into one of the following three categories:

- *Cross-sectional comparison studies*: In a cross-sectional comparison study, the clinician interprets each case first without the assistance of CAD and then, after formally recording his or her findings, interprets the case again while reviewing the CAD results. This specific study design, which uses the same clinician and the same patient as the control, can compare the interpretation without and with CADe as a second reader, with fewer confounding factors and less variability than the historical-control design. However, it requires the recording of both the unaided and aided reading results. Recording of the intermediate unaided readings, which may not be a part of standard practice, may be cumbersome for the reader, and can impact the interpretation process and study conclusions. The cross-sectional design has been used in the evaluation of CADe in

mammography and CT colonography in several prospective studies.[125–130]

- *Historical-control studies*: In a historical-control study, the interpretations of a group of clinicians with CAD over a period of time is compared collectively to the unaided readings of the same or another group of clinicians in a different (usually prior) time interval, or to compare the cancer detection rate and other metrics in a clinic before and after implementation of CAD.[131–133] An advantage of this approach is that it can be implemented directly within routine clinical practice. A disadvantage is that other longitudinal changes that may have occurred between the control period and the study period will confound the results of the performance comparison.
- *Comparison with double reading studies*: In a comparison with double reading, clinicians' performance with CAD is compared with that of double reading, which consists of having two clinicians interpret each case separately.[134,135] This type of studies may be performed either retrospectively in the fashion of case review or prospectively within a large prospective trial.

An important consideration in a prospective CAD reader study is the choice of endpoints. Assessing CAD before, during, or after its introduction into clinical practice may require different study designs and study endpoints.[136] Nishikawa and Pesce argue that evaluation of the impact of CAD using a historical-control study design is fundamentally different from using a cross-sectional study design, especially when the cancer detection rate is chosen as the endpoint which may not correctly measure the clinical impact of CAD.[137] This is because the introduction of CAD affects the number of detectable cancers in the target population so that the target populations in the two time periods are different. Therefore, it is possible that the cancer detection rate would not be substantially different between the two arms of the study even when the CAD actually does lead to earlier, and potentially additional, cancer detections. They suggested that alternative endpoints to assess the effectiveness of CAD, such as changes in size, stage, or nodal status, may be more appropriate in these studies.[137] The results from a recent study by Fenton *et al.*[136] may be an illustration of these effects. The authors reported that the use of CAD was associated with greater rate of detection for ductal carcinoma *in situ* (DCIS), albeit at the expense of an increase in diagnostic workup among women without breast cancer, but no difference in invasive breast cancer incidence. They also reported that CAD was associated with greater likelihood of stage I to II versus III to IV cancer being found among women with invasive cancer. These observations suggest that CAD may have served the purpose of improving detection of cancers at the early stages even though the total cancer detection rate might not change.

7.C.2. Retrospective CAD studies

Prospective testing is ideal for evaluating the utility of CAD in the true clinical setting. However, prospective testing may require an excessively large number of patients when the disease prevalence is low, and a retrospective reader study might serve as an appropriate alternative for some purposes such as initial assessment of the effects of a given CAD system on reader performance. In a retrospective reader study design, cases are collected from a patient archive and are read offline by one or more readers. For low-prevalence diseases, the most common approach for retrospective CAD evaluation is to use an enriched reader study design whereby the population of cases is enriched with patients known to be diseased. This approach has the advantage of substantially reducing the number of cases required for achieving statistically significant results. However, retrospective studies may also impact the behavior of readers compared with clinical practice because readers know that they are participating in a study that does not affect patient management. Readers may also become cognizant of the enrichment relative to their clinical practice and adapt accordingly. Likewise, the case mix can influence the study results. Under these conditions, the results of a retrospective study for each modality (with-CAD and without-CAD) may not generalize to clinical practice. However, it is generally assumed that each modality is affected similarly by the nonclinical study conditions and data set. The without-CAD modality is used to benchmark reader performance allowing it to be compared with the with-CAD performance. Retrospective reader studies are the typical study design used by CAD manufacturers to support the FDA approval of CAD devices.

7.C.3. Performance metrics used in clinical assessment studies

The performance metrics utilized in the reader performance assessment of CAD are the same as those used in standalone performance (see Secs. 5 and 6). In addition, a region-based ROC analysis approach[138,139] has been used in reader studies, but it is not commonly used in standalone testing. In this approach, the image (or patient) is divided into ROIs (e.g., five lobes of the lung). Instead of performing ROC analysis at the patient level, the ROI is treated as the unit of analysis, and each ROI is scored as in conventional ROC analysis regardless of the location of the detected lesion within the ROI. ROI-based performance metrics include the AUC, sensitivity, specificity, and others as described above.

7.D. Reader training for participation in MRMC studies

User training for the appropriate use of CAD in the clinical setting is discussed in a companion paper.[14] In MRMC studies, appropriate reader training on both the use of CAD and on how to participate as a reader is extremely important. CAD training should include at least a description of the function of the CAD system and its average performance, the type of CAD information provided to the user, how CAD would be implemented as part of the clinical workflow, instructions on how to follow the study protocol, on the appropriate use of any study-specific scoring or rating scales, and on the use of

any data entry forms. Training should also include reading a set of example cases, which are not used in the actual reader study, without and with CAD and feedback of typical TPs, FPs, and FNs of the CAD system. In addition, a pilot study is recommended to identify pitfalls in the proposed study protocol or reader training.

8. SAMPLE SIZE ESTIMATION

The sample size used in a CAD assessment study is an important factor that affects the confidence intervals of the observed FOM(s), which, in turn, may influence the conclusions that can be drawn from the study. Sample size refers to the number of cases in CAD assessment, or to the number of cases, the number of readers, or both in a reader study. Sample size estimation is critical in CAD assessment because data set collection and reader studies are resource intensive. Both under- and over-powered studies are undesirable because the former may have an unacceptable risk for type II error (failing to reject the null hypothesis when it is actually false), and the latter may allocate unnecessarily large resources to the study.

A number of research groups have developed sample size estimation methods targeted at diagnostic imaging. Obuchowski[140] conducted a review of methods that consider the following FOMs: sensitivity/specificity pair, sensitivity at a particular FPF, likelihood ratio, full AUC, and partial AUC. She also discussed a method for sample size estimation for MRMC ROC studies.[122,140] More recently, Hillis *et al.* provided a power and sample size estimation method for the DBM method,[141] and unified sample size estimation techniques for the DBM and OR methods.[142] Gallas *et al.*[118] compared the power of different reader study designs in CAD assessment by using a variance expression derived for a nonparametric MRMC AUC estimate. A web applet is available to perform sample size estimation using a variety of variance estimation methods for the AUC.[143] Obuchowski and Hillis provided sample size tables for MRMC CAD studies that account for multiple actual positives using the region-based ROC analysis approach.[144] Chakraborty discussed how methods designed for ROC analysis can be adapted for sample size estimation in location-specific ROC analysis.[145]

Components of variance and effect size are important parameters needed for sample size estimation. A pilot study, conducted before a pivotal study, may be the best way to estimate these parameters. Besides sample size estimation, a pilot study on CAD assessment may provide crucial information on a variety of study design issues, including reading protocol and reader training as discussed previously.

9. CONCLUSION

A number of CAD devices have been approved for clinical use and many more are currently under development. These devices are designed to provide decision support to clinicians in tasks such as risk prediction, disease detection, differential diagnosis, and treatment decisions. At different stages of development, investigators may be interested in assessing different aspects of CAD performance: standalone performance, relative performance compared with other CAD systems, improvement in clinicians' performance in controlled studies compared to that without CAD, and ultimately, improvement in clinical practice. The AAPM CADSC has attempted to identify and provide initial opinions on important components to be considered or to be included as a part of CAD study design in an effort to stimulate further discussion and to help develop a consensus on appropriate approaches for CAD system assessment.

The CADSC identified the following major areas for consideration when assessing CAD: training and test data sets; reference standards; mark-labeling criteria; standalone performance assessment metrics and methodologies; reader performance assessment metrics and methodologies; and study sample size estimation. For each of these areas, we summarized the current state of knowledge, identified practical techniques and methodologies to be followed, provided recommendations that might be useful in real-world situations, and identified some reporting and study design requirements that may be critical. We also discussed areas where further research is needed. We hope that the ideas discussed herein will serve as a framework for further development of structured, unified guidelines for CAD performance assessment and help improve the reliability of reported performance results. Although most of the discussion focuses on the more commonly used lesion detection and diagnosis systems, the principles and basic approaches should serve as a guide for performance assessment of other CAD systems. Proper assessment of the CAD system standalone performance or its impact on the user will lead to a better understanding of its effectiveness and limitations, which, in turn, is expected to stimulate further research and development efforts on CAD technologies, reduce problems due to improper use, and eventually improve the utility and efficacy of CAD in clinical practice.

ACKNOWLEDGMENTS

The authors are grateful to the members and participants of the CADSC who have contributed to the stimulating discussions during many meetings and teleconferences. R.M.S. is supported in part by the Intramural Research Program of the National Institutes of Health, Clinical Center; the views expressed in this paper are the opinions of the authors and do not necessarily represent the views of the National Institutes of Health or the Department of Health and Human Services. S.G.A. and H.Y. receive royalties and licensing fees through the University of Chicago related to CAD. R.M.S. receives patent royalties and research support related to CAD from iCAD. M.T.F. receives funding from Riverain Technologies through Georgetown University Medical Center.

APPENDIX: AAPM CAD SUBCOMMITTEE MEMBERSHIP

Chairs:	Heang-Ping Chan, University of Michigan
	Samuel Armato III, The University of Chicago
Group 1 leader:	Berkman Sahiner, FDA
Group 2 leader:	Nicholas Petrick, FDA
Group 3 leader:	Zhimin Huo, Carestream Health, Inc.
Group 4 leader:	Ronald Summers, NIH Clinical Center

Members and Participants:
Stephen Aylward, Kitware
Alberto Bert, im3d Medical Imaging
Loredana Correale, im3d Medical Imaging
Silvia Delsanto, im3d Medical Imaging
Matthew T. Freedman, Georgetown University
David Fryd, Riverain Medical
Hiroshi Fujita, Gifu University
David Gur, University of Pittsburgh
Lubomir Hadjiiski, University of Michigan
Akira Hasegawa, FUJIFILM Medical Systems
Jeffrey Hoffmeister, iCAD, Inc.
Yulei Jiang, The University of Chicago
Nico Karssemeijer, Radboud University
Jesse Lin, FUJIFILM Medical Systems
Shih-Chung Ben Lo, Georgetown University
Joseph Lo, Duke University
Mia Markey, University of Texas at Austin
Julian Marshall, Hologic, Inc.
Michael McNitt-Gray, UCLA
Patricia Milbank
Lia Morra, im3d Medical Imaging
Sophie Paquerault
Vikas Raykar, Siemens Medical
Anthony Reeves, Cornell University
Marcos Salganicoff, Siemens Medical
Frank Samuelson, FDA
Eric Silfen, Philips Healthcare
Georgia Tourassi, Duke University (Oak Ridge National Laboratory)
Stephen Vastagh, MITA
Hiroyuki Yoshida, Massachusetts General Hospital/Harvard University
Bin Zheng, University of Pittsburgh
Chuan Zhou, University of Michigan

[a]An opinion paper from the American Association of Physicists in Medicine (AAPM) Computer Aided Detection in Diagnostic Imaging subcommittee (CADSC).
[b]These two co-authors contributed equally to this work.
[c]Author to whom correspondence should be addressed. Electronic mail: chanhp@umich.edu; Telephone: 734-936-4357.

[1]L. B. Lusted, "Logical analysis in roentgen diagnosis - Memorial fund lecture," Radiology **74**, 178–193 (1960).
[2]W. J. Tuddenham, "Visual search, image organization, and reader error in roentgen diagnosis – Studies of the psychophysiology of roentgen image perception - Memorial fund lecture," Radiology **78**, 694–704 (1962).
[3]H. L. Kundel and G. Revesz, "Lesion conspicuity, structured noise, and film reader error," Am. J. Roentgenol. **126**, 1233–1238 (1976).
[4]K. S. Berbaum, E. A. Franken, D. D. Dorfman, S. A. Rooholamini, M. H. Kathol, T. J. Barloon, F. M. Behlke, Y. Sato, C. H. Lu, G. Y. Elkhoury, F. W. Flickinger, and W. J. Montgomery, "Satisfaction of Search in diagnostic-radiology," Invest. Radiol. **25**, 133–140 (1990).
[5]D. L. Renfrew, E. A. Franken, K. S. Berbaum, F. H. Weigelt, and M. M. Abuyousef, "Error in radiology – Classification and lessons in 182 cases presented at a problem case conference," Radiology **183**, 145–150 (1992).
[6]K. Doi, "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential," Comput. Med. Imaging Graph. **31**, 198–211 (2007).
[7]D. Bielen and G. Kiss, "Computer-aided detection for CT colonography: Update 2007," Abdom. Imaging **32**, 571–581 (2007).
[8]M. L. Giger, H. P. Chan, and J. Boone, "Anniversary paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM," Med. Phys. **35**, 5799–5820 (2008).
[9]H. P. Chan, L. Hadjiiski, C. Zhou, and B. Sahiner, "Computer-aided diagnosis of lung cancer and pulmonary embolism in computed tomography - A review," Acad. Radiol. **15**, 535–555 (2008).
[10]B. van Ginneken, L. Hogeweg, and M. Prokop, "Computer-aided diagnosis in chest radiography: Beyond nodules," Eur. J. Radiol. **72**, 226–230 (2009).
[11]M. Elter and A. Horsch, "CADx of mammographic masses and clustered microcalcifications: A review," Med. Phys. **36**, 2052–2068 (2009).
[12]H. D. Cheng, J. Shan, W. Ju, Y. H. Guo, and L. Zhang, "Automated breast cancer detection and classification using ultrasound images: A survey," Pattern Recognit. **43**, 299–317 (2010).
[13]A. Oliver, J. Freixenet, J. Marti, E. Perez, J. Pont, E. R. E. Denton, and R. Zwiggelaar, "A review of automatic mass detection and segmentation in mammographic images," Med. Image Anal. **14**, 87–110 (2010).
[14]Z. Huo, R. M. Summers, S. Paquerault, J. Lo, J. Hoffmeister, S. G. Armato III, M. T. Freedman, J. Lin, S.-C. B. Lo, N. Petrick, B. Sahiner, D. Fryd, H. Yoshida, and H.-P. Chan, "Quality assurance and training procedures for computer-aided detection and diagnosis systems in clinical use," Med. Phys. **40**, 077001 (13pp.) (2013).
[15]B. Efron, "Estimating the error rate of a prediction rule - Improvement on cross-validation," J. Am. Stat. Assoc. **78**, 316–331 (1983).
[16]T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning* (Springer-Verlag, New York, 2001).
[17]C. M. Bishop, *Neural Networks for Pattern Recognition* (Clarendon Press, Oxford, 1995).
[18]L. E. Dodd, R. F. Wagner, S. G. Armato, M. F. McNitt-Gray, S. Beiden, H. P. Chan, D. Gur, G. McLennan, C. E. Metz, N. Petrick, B. Sahiner, J. Sayre, and R. Lung Image Database Consortium, "Assessment methodologies and statistical issues for computer-aided diagnosis of lung nodules in computed tomography: Contemporary research topics relevant to the lung image database consortium," Acad. Radiol. **11**, 462–475 (2004).
[19]M. T. Madsen, K. S. Berbaum, A. N. Ellingson, B. H. Thompson, B. F. Mullan, and R. T. Caldwell, "A new software tool for removing, storing, and adding abnormalities to medical images for perception research studies," Acad. Radiol. **13**, 305–312 (2006).
[20]X. Li, E. Samei, D. M. Delong, R. P. Jones, A. M. Gaca, C. L. Hollingsworth, C. M. Maxfield, C. W. T. Carrico, and D. P. Frush, "Three-dimensional simulation of lung nodules for paediatric multidetector array CT," Br. J. Radiol. **82**, 401–411 (2009).
[21]H. L. Kundel, "Disease prevalence and radiological decision-making," Invest. Radiol. **17**, 107–109 (1982).
[22]R. F. Wagner, C. A. Beam, and S. V. Beiden, "Reader variability in mammography and its implications for expected utility over the population of readers and cases," Med. Decis. Making **24**, 561–572 (2004).
[23]D. Gur, A. I. Bandos, C. R. Fuhrman, A. H. Klym, J. L. King, and H. E. Rockette, "The prevalence effect in a laboratory environment: Changing the confidence ratings," Acad. Radiol. **14**, 49–53 (2007).
[24]M. S. Pepe, *The Statistical Evaluation of Medical Tests for Classification and Prediction* (Oxford University Press, New York, 2003).
[25]C. B. Begg and R. A. Greenes, "Assessment of diagnostic tests when disease verification is subject to selection bias," Biometrics **39**, 207–215 (1983).
[26]R. F. Wagner, C. E. Metz, and G. Campbell, "Assessment of medical imaging systems and computer aids: A tutorial review," Acad. Radiol. **14**, 723–748 (2007).
[27]B. Zheng, M. A. Ganott, C. A. Britton, C. M. Hakim, L. A. Hardesty, T. S. Chang, H. E. Rockette, and D. Gur, "Soft-copy mammographic readings with different computer-assisted detection cuing environments: Preliminary findings," Radiology **221**, 633–640 (2001).
[28]M. Heath, K. Bowyer, D. Kopans, R. Moore, and P. Kegelmeyer, "The digital database for screening mammography," in *Digital Mammography; IWDM 2000*, edited by M. J. Yaffe (Medical Physics, Toronto, Canada, 2001), pp. 457–460.

[29]J. Shiraishi, S. Katsuragawa, J. Ikezoe, T. Matsumoto, T. Kobayashi, K. Komatsu, M. Matsui, H. Fujita, Y. Kodera, and K. Doi, "Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules," Am. J. Roentgenol. **174**, 71–74 (2000).
[30]S. G. Armato, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. S. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman, E. A. Kazerooni, H. MacMahon, E. J. R. van Beek, D. Yankelevitz, A. M. Biancardi, P. H. Bland, M. S. Brown, R. M. Engelmann, G. E. Laderach, D. Max, R. C. Pais, D. P. Y. Qing, R. Y. Roberts, A. R. Smith, A. Starkey, P. Batra, P. Caligiuri, A. Farooqi, G. W. Gladish, C. M. Jude, R. F. Munden, I. Petkovska, L. E. Quint, L. H. Schwartz, B. Sundaram, L. E. Dodd, C. Fenimore, D. Gur, N. Petrick, J. Freymann, J. Kirby, B. Hughes, A. V. Casteele, S. Gupte, M. Sallam, M. D. Heath, M. H. Kuhn, E. Dharaiya, R. Burns, D. S. Fryd, M. Salganicoff, V. Anand, U. Shreter, S. Vastagh, B. Y. Croft, and L. P. Clarke, "The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans," Med. Phys. **38**, 915–931 (2011).
[31]S. S. Raab, D. M. Grzybicki, J. E. Janosky, R. J. Zarbo, F. A. Meier, C. Jensen, and S. J. Geyer, "Clinical impact and frequency of anatomic pathology errors in cancer diagnoses," Cancer **104**, 2205–2213 (2005).
[32]D. P. Miller, K. F. O'Shaughnessy, S. A. Wood, and R. A. Castellino, "Gold standards and expert panels: A pulmonary nodule case study with challenges and solutions," Proc. SPIE **5372**, 173–184 (2004).
[33]M. Das, G. Muhlenbruch, A. H. Mahnken, T. G. Flohr, L. Gundel, S. Stanzel, T. Kraus, R. W. Gunther, and J. E. Wildberger, "Small pulmonary nodules: Effect of two computer-aided detection systems on radiologist performance," Radiology **241**, 564–571 (2006).
[34]S. Buhmann, P. Herzog, J. Liang, M. Wolf, M. Salganicoff, C. Kirchhoff, M. Reiser, and C. H. Becker, "Clinical evaluation of a computer-aided diagnosis (CAD) prototype for the detection of pulmonary embolism," Acad. Radiol. **14**, 651–658 (2007).
[35]A. M. Biancardi, A. C. Jirapatnakul, and A. P. Reeves, "A comparison of ground truth estimation methods," Int. J. Comput. Assist. Radiol. Surg. **5**, 295–305 (2010).
[36]K. R. Choudhury, D. S. Paik, C. A. Yi, S. Napel, J. Roos, and G. D. Rubin, "Assessing operating characteristics of CAD algorithms in the absence of a gold standard," Med. Phys. **37**, 1788–1795 (2010).
[37]S. G. Armato, R. Y. Roberts, M. Kocherginsky, D. R. Aberle, E. A. Kazerooni, H. MacMahon, E. J. R. van Beek, D. Yankelevitz, G. McLennan, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, P. Caligiuri, L. E. Quint, B. Sundaram, B. Y. Croft, and L. P. Clarke, "Assessment of radiologist performance in the detection of lung nodules: Dependence on the definition of 'Truth'," Acad. Radiol. **16**, 28–38 (2009).
[38]A. M. R. Schilham, B. van Ginneken, and M. Loog, "A computer-aided diagnosis system for detection of lung nodules in chest radiographs with an evaluation on a public database," Med. Image Anal. **10**, 247–258 (2006).
[39]R. M. Summers, L. R. Handwerker, P. J. Pickhardt, R. L. Van Uitert, K. K. Deshpande, S. Yeshwant, J. Yao, and M. Franaszek, "Performance of a previously validated CT colonography computer-aided detection system in a new patient population," Am. J. Roentgenol. **191**, 168–174 (2008).
[40]U. Bick, M. L. Giger, R. A. Schmidt, R. M. Nishikawa, D. E. Wolverton, and K. Doi, "Computer-aided breast cancer detection in screening mammography," in *Digital Mammography*, edited by A. G. Gale, S. M. Astley, D. R. Dance, and A. Y. Cairns (Elsevier, Amsterdam, 1996).
[41]K. G. Kim, J. M. Goo, J. H. Kim, H. J. Lee, B. G. Min, K. T. Bae, and J.-G. Im, "Computer-aided diagnosis of localized ground-glass opacity in the lung at CT: Initial experience," Radiology **237**, 657–661 (2005).
[42]N. Petrick, B. Sahiner, H. P. Chan, M. A. Helvie, S. Paquerault, and L. M. Hadjiiski, "Breast cancer detection: Evaluation of a mass-detection algorithm for computer-aided diagnosis - Experience in 263 patients," Radiology **224**, 217–224 (2002).
[43]H. D. Li, M. Kallergi, L. P. Clarke, V. K. Jain, and R. A. Clark, "Markov random field for tumor detection in digital mammography," IEEE Trans. Med. Imaging **14**, 565–576 (1995).
[44]X. W. Xu, K. Doi, T. Kobayashi, H. MacMahon, and M. L. Giger, "Development of an improved CAD scheme for automated detection of lung nodules in digital chest images," Med. Phys. **24**, 1395–1403 (1997).
[45]B. Keserci and H. Yoshida, "Computerized detection of pulmonary nodules in chest radiographs based on morphological features and wavelet snake model," Med. Image Anal. **6**, 431–447 (2002).
[46]H. Yoshida, Y. Masutani, P. MacEneaney, D. T. Rubin, and A. H. Dachman, "Computerized detection of colonic polyps at CT colonography on the basis of volumetric features: Pilot study," Radiology **222**, 327–336 (2002).
[47]H.-P. Chan, S. C. B. Lo, B. Sahiner, K. L. Lam, and M. A. Helvie, "Computer-aided detection of mammographic microcalcifications: Pattern recognition with an artificial neural network," Med. Phys. **22**, 1555–1567 (1995).
[48]A. A. Enquobahrie, A. P. Reeves, D. F. Yankelevitz, and C. I. Henschke, "Automated detection of small pulmonary nodules in whole lung CT scans," Acad. Radiol. **14**, 579–593 (2007).
[49]D. S. Paik, C. F. Beaulieu, G. D. Rubin, B. Acar, R. B. Jeffrey, Jr., J. Yee, J. Dey, and S. Napel, "Surface normal overlap: A computer-aided detection algorithm with application to colonic polyps and lung nodules in helical CT," IEEE Trans. Med. Imaging **23**, 661–675 (2004).
[50]C. S. White, T. Flukinger, J. Jeudy, and J. J. Chen, "Use of a computer-aided detection system to detect missed lung cancer at chest radiography," Radiology **252**, 273–281 (2009).
[51]J. Dehmeshki, S. Halligan, S. A. Taylor, M. E. Roddie, J. McQuillan, L. Honeyfield, and H. Amin, "Computer assisted detection software for CT colonography: Effect of sphericity filter on performance characteristics for patients with and without fecal tagging," Eur. Radiol. **17**, 662–668 (2007).
[52]Q. Li, F. Li and K. Doi, "Computerized detection of lung nodules in thin-section CT images by use of selective enhancement filters and an automated rule-based classifier," Acad. Radiol. **15**, 165–175 (2008).
[53]S. G. Armato, "Image annotation for conveying automated lung nodule detection results to radiologists," Acad. Radiol. **10**, 1000–1007 (2003).
[54]S. A. Taylor, J. Brittenden, J. Lenton, H. Lambie, A. Goldstone, P. N. Wylie, D. Tolan, D. Burling, L. Honeyfield, P. Bassett, and S. Halligan, "Influence of computer-aided detection false-positives on reader performance and diagnostic confidence for CT colonography," Am. J. Roentgenol. **192**, 1682–1689 (2009).
[55]M. Kallergi, G. M. Carney, and J. Garviria, "Evaluating the performance of detection algorithms in digital mammography," Med. Phys. **26**, 267–275 (1999).
[56]M. H. Zweig and G. Campbell, "Receiver-operating characteristic (ROC) plots - A fundamental evaluation tool in clinical medicine," Clin. Chem. **39**, 561–577 (1993).
[57]D. D. Dorfman and E. Alf, Jr., "Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data," J. Math. Psychol. **6**, 487–496 (1969).
[58]D. D. Dorfman, K. S. Berbaum, C. E. Metz, R. V. Lenth, J. A. Hanley, and H. AbuDagga, "Proper receiver operating characteristic analysis: The bigamma model," Acad. Radiol. **4**, 138–149 (1997).
[59]C. E. Metz and X. Pan, "Proper'' binormal ROC curves: Theory and maximum-likelihood estimation," J. Math. Psychol. **43**, 1–33 (1999).
[60]D. D. Dorfman and K. S. Berbaum, "A contaminated binormal model for ROC data - Part II. A formal model," Acad. Radiol. **7**, 427–437 (2000).
[61]Y. L. Jiang, C. E. Metz, and R. M. Nishikawa, "A receiver operating characteristic partial area index for highly sensitive diagnostic tests," Radiology **201**, 745–750 (1996).
[62]E. R. Delong, D. M. Delong, and D. I. Clarkepearson, "Comparing the areas under 2 or more correlated receiver operating characteristic curves - A nonparametric approach," Biometrics **44**, 837–845 (1988).
[63]J. P. Egan, *Signal Detection Theory and ROC Analysis* (Academic, New York, 1975).
[64]C. E. Metz, "Basic principles of ROC analysis," Semin. Nucl. Med. **8**, 283–298 (1978).
[65]J. Hanley and B. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve," Radiology **143**, 29–36 (1982).
[66]C. E. Metz, "Some practical issues of experimental design and data analysis in radiological ROC studies," Invest. Radiol. **24**, 234–245 (1989).
[67]J. A. Swets, *Signal Detection Theory and ROC Analysis in Psychology and Diagnostics: Collected Papers* (Lawrence Erlbaum Associates, NJ, 1996).
[68]L. E. Dodd and M. S. Pepe, "Partial AUC estimation and regression," Biometrics **59**, 614–623 (2003).
[69]N. Obuchowski, "Receiver operating characteristic curves and their use in radiology," Radiology **229**, 3–8 (2003).
[70]K. H. Zou, A. J. O'Malley, and L. Mauri, "Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models," Circulation **115**, 654–657 (2007).

[71]ICRU, "Receiver operating characteristic analysis in medical imaging," Report No. 79 (International Commission of Radiation Units and Measurements, Bethesda, MD, 2008).
[72]X. He and E. Frey, "ROC, LROC, FROC, AFROC: An alphabet soup," J. Am. Coll. Radiol. **6**, 652–655 (2009).
[73]See: http://www.bio.ri.ccf.org/html/rocanalysis.html for Cleveland Clinic ROC Software.
[74]See: http://metz-roc.uchicago.edu/ for University of Chicago LABMRMC and CORROC Software.
[75]See: http://perception.radiology.uiowa.edu/ for University of Iowa MRMC Software.
[76]N. A. Obuchowski, "Fundamentals of clinical research for radiologists - ROC analysis," Am. J. Roentgenol. **184**, 364–372 (2005).
[77]S. J. Starr, C. E. Metz, L. B. Lusted, and D. J. Goodenough, "Visual detection and localization of radiographic images," Radiology **116**, 533–538 (1975).
[78]P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons, "A free response approach to the measurement and characterization of radiographic observer performance," Proc. SPIE **0127**, 124–135 (1977).
[79]R. G. Swensson, "Unified measurement of observer performance in detection and localizing target objects on images," Med. Phys. **23**, 1709–1724 (1996).
[80]D. C. Edwards, M. A. Kupinski, C. E. Metz, and R. M. Nishikawa, "Maximum likelihood fitting of FROC curves under an initial-detection-and-candidate-analysis model," Med. Phys. **29**, 2861–2870 (2002).
[81]D. P. Chakraborty, "A search model and figure of merit for observer data acquired according to the free-response paradigm," Phys. Med. Biol. **51**, 3449–3462 (2006).
[82]H. J. Yoon, B. Zheng, B. Sahiner, and D. P. Chakraborty, "Evaluating computer-aided detection algorithms," Med. Phys. **34**, 2024–2038 (2007).
[83]See: http://www.devchakraborty.com/ for JAFROC software.
[84]D. P. Chakraborty and K. S. Berbaum, "Observer studies involving detection and localization: Modeling, analysis, and validation," Med. Phys. **31**, 2313–2330 (2004).
[85]D. P. Chakraborty, "Analysis of location specific observer performance data: Validated extensions of the jackknife free-response (JAFROC) method," Acad. Radiol. **13**, 1187–1193 (2006).
[86]A. I. Bandos, H. E. Rockette, T. Song, and D. Gur, "Area under the free-response ROC curve (FROC) and a related summary index," Biometrics **65**, 247–256 (2009).
[87]H. Bornefalk and A. B. Hermansson, "On the comparison of FROC curves in mammography CAD systems," Med. Phys. **32**, 412–417 (2005).
[88]F. W. Samuelson and N. Petrick, "Comparing image detection algorithms using resampling," in *3rd IEEE International Symposium on Biomedical Imaging: Macro to Nano* (2006), Vols. 1–3, pp. 1312–1315.
[89]L. M. Popescu, "Nonparametric signal detectability evaluation using an exponential transformation of the FROC curve," Med. Phys. **38**, 5690–5702 (2011).
[90]D. P. Chakraborty, "Validation and statistical power comparison of methods for analyzing free-response observer performance studies," Acad. Radiol. **15**, 1554–1566 (2008).
[91]L. D. Brown, T. T. Cai, and A. DasGupta, "Interval estimation for a binomial proportion," Stat. Sci. **16**, 101–117 (2001).
[92]L. D. Brown, T. T. Cai, A. DasGupta, A. Agresti, B. A. Coull, G. Casella, C. Corcoran, C. Mehta, M. Ghosh, and T. J. Santner, "Interval estimation for a binomial proportion - Comment - Rejoinder," Stat. Sci. **16**, 101–133 (2001).
[93]L. D. Brown, T. T. Cai, and A. DasGupta, "Interval estimation in exponential families," Stat. Sin. **13**, 19–49 (2003).
[94]A. Agresti and B. A. Coull, "Approximate is better than "exact" for interval estimation of binomial proportions," Am. Stat. **52**, 119–126 (1998).
[95]P. C. Bunch, J. F. Hamilton, G. K. Sanderson, and A. H. Simmons, "A free response approach to the measurement and characterization of radiographic observer performance," J. Appl. Photogr. Eng. **4**, 166–171 (1978).
[96]D. P. Chakraborty, "Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data," Med. Phys. **16**, 561–568 (1989).
[97]D. P. Chakraborty and L. H. L. Winter, "Free-response methodology: Alternate analysis and a new observer-performance experiment," Radiology **174**, 873–881 (1990).
[98]J. M. Neuhaus and J. D. Kalbfleisch, "Between- and within-cluster covariate effects in the analysis of clustered data," Biometrics **54**, 638–645 (1998).
[99]J. N. K. Rao and A. J. Scott, "A simple method for the analysis of clustered binary data," Biometrics **48**, 577–585 (1992).
[100]B. Efron and R. Tibshirani, "Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy," Stat. Sci. **1**, 54–75 (1986).
[101]J. J. Nappi and K. Nagata, "Sources of false positives in computer-assisted CT colonography," Abdom. Imaging **36**, 153–164 (2011).
[102]V. S. Koshkin, J. L. Hinshaw, K. Wroblewski, and A. H. Dachman, "CAD-associated reader error in CT colonography," Acad. Radiol. **19**, 801–810 (2012).
[103]D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," Mach. Learn. **45**, 171–186 (2001).
[104]C. Ferri, J. Hernandez-Orallo, and M. A. Salido, "Volume under the ROC surface for multi-class problems," in *Machine Learning: ECML*, edited by N. Lavrac, D. Gamberger, H. Blockeel, and L. Todorovski (Cavtat-Dubrovnik, Croatia, Springer, 2003), Vol. 2837, pp. 108–120.
[105]B. K. Scurfield, "Multiple-event forced-choice tasks in the theory of signal detectability," J. Math. Psychol. **40**, 253–269 (1996).
[106]D. Mossman, "Three-way ROCs," Med. Decis. Making **19**, 78–89 (1999).
[107]D. C. Edwards, C. E. Metz, and M. A. Kupinski, "Ideal observers and optimal ROC hypersurfaces in N-class classification," IEEE Trans. Med. Imaging **23**, 891–895 (2004).
[108]X. He, C. E. Metz, B. M. W. Tsui, J. M. Links, and E. C. Frey, "Three-class ROC analysis – A decision theoreric approach under the ideal observer framework," IEEE Trans. Med. Imaging **25**, 571–581 (2006).
[109]B. Sahiner, H.-P. Chan, and L. Hadjiiski, "Performance analysis of 3-class classifiers: Properties of the 3D ROC surface and the normalized volume under the surface for the ideal observer," IEEE Trans. Med. Imaging **27**, 215–227 (2008).
[110]X. He and E. C. Frey, "The meaning and use of the volume under a three-class ROC surface (VUS)," IEEE Trans. Med. Imaging **27**, 577–588 (2008).
[111]C. Beam, P. Layde, and D. Sullivan, "Variability in the interpretation of screening mammograms by US Radiologists," Arch. Intern Med. **156**, 209–213 (1996).
[112]N. Petrick, M. Haider, R. M. Summers, S. C. Yeshwant, L. Brown, E. M. Iuliano, A. Louie, J. R. Choi, and P. J. Pickhardt, "CT colonography with computer-aided detection as a second reader: Observer performance study," Radiology **246**, 148–156 (2008).
[113]S. A. Taylor, S. C. Charman, P. Lefere, E. G. McFarland, E. K. Paulson, J. Yee, R. Aslam, J. M. Barlow, A. Gupta, D. H. Kim, C. M. Miller, S. Halligan, S. A. Taylor, S. C. Charman, P. Lefere, E. G. McFarland, E. K. Paulson, J. Yee, R. Aslam, J. M. Barlow, A. Gupta, D. H. Kim, C. M. Miller, and S. Halligan, "CT colonography: Investigation of the optimum reader paradigm by using computer-aided detection software," Radiology **246**, 463–471 (2008).
[114]H. P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu, and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms. The potential of computer-aided diagnosis," Invest. Radiol. **25**, 1102–1110 (1990).
[115]S. H. Taplin, C. M. Rutter, and C. D. Lehman, "Testing the effect of computer-assisted detection on interpretive performance in screening mammography.[see comment]," AJR, Am. J. Roentgenol. **187**, 1475–1482 (2006).
[116]E. A. Krupinski, "Perceptual enhancement of pulmonary nodule recognition in chest radiographs," Proc. SPIE **2166**, 59–65 (1994).
[117]N. A. Obuchowski, "Reducing the number of reader interpretations in MRMC studies," Acad. Radiol. **16**, 209–217 (2009).
[118]B. D. Gallas and D. G. Brown, "Reader studies for validation of CAD systems," Neural Netw. **21**, 387–397 (2008) [erratum in Neural Netw. **21**(4), 698 (2008)].
[119]N. A. Obuchowski, S. V. Beiden, K. S. Berbaum, S. L. Hillis, H. Ishwaran, H. H. Song, and R. F. Wagner, "Multireader, multicase receiver operating characteristic analysis: An empirical comparison of five methods," Acad. Radiol. **11**, 980–995 (2004).
[120]D. D. Dorfman, K. S. Berbaum, and C. E. Metz, "Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method," Invest. Radiol. **27**, 723–731 (1992).
[121]N. A. Obuchowski and H. E. Rockette, "Hypothesis testing of diagnostic accuracy for multiple readers and multiple tests an anova approach with dependent observations," Commun. Stat. Simul. Comput. **24**, 285–308 (1995).

[122]N. A. Obuchowski, "Multireader, multimodality receiver operating characteristic curve studies: Hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations," Acad. Radiol. **2**(1), S22–S29 (1995), discussion S57–S64.
[123]B. D. Gallas, "One-shot estimate of MRMC variance: AUC," Acad. Radiol. **13**, 353–362 (2006).
[124]B. D. Gallas, G. A. Pennello, and K. J. Myers, "Multireader multicase variance analysis for binary data," J. Opt. Soc. Am. A **24**, B70–B80 (2007).
[125]T. W. Freer and M. J. Ulissey, "Screening mammography with computer-aided detection: Prospective study of 12,860 patients in a community breast center," Radiology **220**, 781–786 (2001).
[126]J. C. Dean and C. C. Ilvento, "Improved cancer detection using computer-aided detection with diagnostic and screening mammography: Prospective study of 104 cancers," AJR, Am. J. Roentgenol. **187**, 20–28 (2006).
[127]M. A. Helvie, L. Hadjiiski, E. Makariou, H. P. Chan, N. Petrick, B. Sahiner, S. C. B. Lo, M. Freedman, D. Adler, J. Bailey, C. Blane, D. Hoff, K. Hunt, L. Joynt, K. Klein, C. Paramagul, S. K. Patterson, and M. A. Roubidoux, "Sensitivity of noncommercial computer-aided detection system for mammographic breast cancer detection: Pilot clinical trial," Radiology **231**, 208–214 (2004).
[128]R. L. Birdwell, P. Bandodkar, and D. M. Ikeda, "Computer-aided detection with screening mammography in a university hospital setting," Radiology **236**, 451–457 (2005).
[129]M. J. Morton, D. H. Whaley, K. R. Brandt, and K. K. Amrami, "Screening mammograms: Interpretation with computer-aided detection - Prospective evaluation," Radiology **239**, 375–383 (2006).
[130]D. Regge, P. Monica, G. Galatola, C. Laudi, A. Zambon, L. Correale, R. Asnaghi, B. Barbaro, C. Borghi, D. Campanella, M. Cassinis, R. Ferrari, A. Ferraris, R. Golfieri, C. Hassan, F. Iafrate, G. Iussich, A. Laghi, R. Massara, E. Neri, L. Sali, S. Venturini, and G. Gandini, "Efficacy of computer-aided detection as a second reader for 6-9-mm lesions at CT colonography: Multicenter prospective trial," Radiology **266**, 168–176 (2013).
[131]D. Gur, J. H. Sumkin, H. E. Rockette, M. Ganott, C. Hakim, L. Hardesty, W. R. Poller, R. Shah, and L. Wallace, "Changes in breast cancer detection and mammography recall rates after the introduction of a computer-aided detection system," J. Natl. Cancer Inst. **96**, 185–190 (2004).
[132]M. Gromet, "Comparison of computer-aided detection to double reading of screening mammograms: Review of 231,221 mammograms," Am. J. Roentgenol. **190**, 854–859 (2008).
[133]J. J. Fenton, L. Abraham, S. H. Taplin, B. M. Geller, P. A. Carney, C. D'Orsi, J. G. Elmore, W. E. Barlow, and C. Breast Cancer Surveillance, "Effectiveness of computer-aided detection in community mammography practice," J. Natl. Cancer Inst. **103**, 1152–1161 (2011).
[134]D. Georgian-Smith, R. H. Moore, E. Halpern, E. D. Yeh, E. A. Rafferty, H. A. D'Alessandro, M. Staffa, D. A. Hall, K. A. McCarthy, and D. B. Kopans, "Blinded comparison of computer-aided detection with human second reading in screening mammography," Am. J. Roentgenol. **189**, 1135–1141 (2007).
[135]F. J. Gilbert, S. M. Astley, M. G. Gillan, O. F. Agbaje, M. G. Wallis, J. James, C. R. Boggis, S. W. Duffy, C. I. Group, F. J. Gilbert, S. M. Astley, M. G. C. Gillan, O. F. Agbaje, M. G. Wallis, J. James, C. R. M. Boggis, and S. W. Duffy, "Single reading with computer-aided detection for screening mammography [see comment]," N. Engl. J. Med. **359**, 1675–1684 (2008).
[136]J. J. Fenton, G. Xing, J. G. Elmore, H. Bang, S. L. Chen, K. K. Lindfors, and L.-M. Baldwin, "Short-term outcomes of screening mammography using computer-aided detection. A population-based study of medicare enrollees," Ann. Intern Med. **158**, 580–587 (2013).
[137]R. M. Nishikawa, and L. L. Pesce, "Computer-aided detection evaluation methods are not created equal," Radiology **251**, 634–636 (2009).
[138]N. A. Obuchowski, M. L. Lieber, and K. A. Powell, "Data analysis for detection and localization of multiple abnormalities with application to mammography," Acad. Radiol. **7**, 516–525 (2000).
[139]C. M. Rutter, "Bootstrap estimation of diagnostic accuracy with patient-clustered data," Acad. Radiol. **7**, 413–419 (2000).
[140]N. A. Obuchowski, "Sample size calculations in studies of test accuracy," Stat. Methods Med. Res. **7**, 371–392 (1998).
[141]S. L. Hillis and K. S. Berbaum, "Power estimation for the Dorfman-Berbaum-Metz method," Acad. Radiol. **11**, 1260–1273 (2004).
[142]S. L. Hillis, N. A. Obuchowski, and K. S. Berbaum, "Power estimation for multireader ROC methods: An updated and unified approach," Acad. Radiol. **18**, 129–142 (2011).
[143]See: http://js.cx/~xin/mrmc.html for iMRMC.
[144]N. A. Obuchowski and S. L. Hillis, "Sample size tables for computer-aided detection studies," Am. J. Roentgenol. **197**, W821–W828 (2011).
[145]D. P. Chakraborty, "New developments in observer performance methodology in medical imaging," Semin. Nucl. Med. **41**, 401–418 (2011).