Quality assurance and training procedures for computer-aided detection and diagnosis systems in clinical use[a]

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

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

Sophie Paquerault[c]
Center for Devices and Radiological Health, U.S. Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, Maryland 20993

Joseph Lo
Duke University Medical Center, 2424 Erwin Road Suite 302, Durham, North Carolina 27705

Jeffrey Hoffmeister
iCAD, Inc., 98 Spitbrook Road, Suite 100, Nashua, New Hampshire 03062

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

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

Jesse Lin
FUJIFILM Medical Systems USA, Inc., 419 West Avenue, Stamford, Connecticut 06902

Shih-Chung Ben Lo
Georgetown University, 3900 Reservoir Road, Northwest, Washington, DC 20057

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

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

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

Heang-Ping Chan[d]
Department of Radiology, The University of Michigan, 1500 East Medical Center Drive, MIB C477, Ann Arbor, Michigan 48109-5842

(Received 26 January 2013; revised 25 April 2013; accepted for publication 7 May 2013; published 11 June 2013)

Computer-aided detection/diagnosis (CAD) is increasingly used for decision support by clinicians for detection and interpretation of diseases. However, there are no quality assurance (QA) requirements for CAD in clinical use at present. QA of CAD is important so that end users can be made aware of changes in CAD performance both due to intentional or unintentional causes. In addition, end-user training is critical to prevent improper use of CAD, which could potentially result in lower overall clinical performance. Research on QA of CAD and user training are limited to date. The purpose of this paper is to bring attention to these issues, inform the readers of the opinions of the members of the American Association of Physicists in Medicine (AAPM) CAD subcommittee, and thus stimulate further discussion in the CAD community on these topics. The recommendations in this paper are intended to be work items for AAPM task groups that will be formed to address QA and user training issues on CAD in the future. The work items may serve as a framework for the discussion and eventual design of detailed QA and training procedures for physicists and users of CAD. Some of the recommendations are considered by the subcommittee to be reasonably easy and practical and can be implemented immediately by the end users; others are considered to be "best practice" approaches, which may require significant effort, additional tools, and proper training to implement. The eventual standardization of the requirements of QA procedures for CAD will have to be determined through consensus from members of the CAD community, and user training may require support of professional societies. It is expected that high-quality CAD and proper use of CAD could allow these systems to achieve their true potential, thus benefiting both the patients and the clinicians, and may bring about more widespread clinical use of CAD for many other diseases and applications. It is hoped that the awareness of the need for appropriate CAD QA and user training will stimulate new ideas and approaches for implementing such procedures efficiently and effectively as well as funding opportunities to fulfill such critical efforts. © *2013 American Association of Physicists in Medicine*. [http://dx.doi.org/10.1118/1.4807642]

Key words: computer-aided diagnosis, computer-aided detection, quality assurance, user training

I. INTRODUCTION

For over a decade, computer-aided detection/diagnosis (CAD) has become a part of routine clinical practice. CAD is increasingly used for decision support by healthcare professionals for detection and interpretation of diseases occurring in various organs of the body.[1] Initial clinical applications of CAD included detection of cancer in breast, lung, and colon examinations. CAD is also rapidly expanding to provide additional functionalities such as estimates of the likelihood of a lesion being malignant or the risk of developing a disease, quantitative measurement tools (e.g., automated segmentation of lesions, size measurement, dynamic flow information), treatment planning, treatment response monitoring, and patient outcome prediction by image-based biomarkers alone or in combination with other clinical information. The acceptance of CAD by clinicians would lead to further improvement in the current CAD systems and opportunities for research and development of new applications.

CAD products can be accepted in clinical use only if they can provide clinical benefit that outweighs the potential associated risks. There are many aspects of CAD that are important to meet this goal. All along the lifecycle of a CAD system (i.e., initiation, concept development, planning, requirements, design, development, verification and validation, implementation, operations, and maintenance), a certain degree of quality assessment is already performed mainly by device manufacturers and by interactions between manufacturers and end users (i.e., healthcare professionals), academia, and government entities. However, there is only limited training offered to the end users and limited quality assurance (QA) testing performed on the system. For example, no requirements or procedures have been developed to assure and analyze the quality of CAD and the use of CAD over time at clinical sites. Therefore, the end users may not be alerted to any unexpected changes in CAD performance and do not have the ability to easily identify any potential issue that may arise.

There are two crucial aspects to consider with CAD:

- **Ensure that CAD products function correctly** when installed at a clinical site and continue to function correctly during clinical use. While these computer-based techniques are typically reliable, consistent, and accurate, inadvertent corruption of these systems or significant changes in a site's imaging systems or differences in patient population may result in a decrease in CAD performance that could substantially change the originally established risk-benefit profile. Therefore, it is important that QA procedures for CAD products be developed, implemented, advocated, and potentially required/enforced at clinical sites so that the end users or QA personnel can routinely check CAD performance and identify any significant changes in performance.
- **Monitor how these CAD products are being used by end users**, particularly clinicians. Even though CAD products are recommended to be used in a certain manner (e.g., second read), the end users may opt for a different use that may be less effective or inappropriate and may change the risk-benefit profile. Studies have shown that training end users on the appropriate use of CAD is an important factor that may limit the potential change of risk-benefit profile of the system.[2]

In late 2007, the American Association of Physicists in Medicine (AAPM) officially created a Computer Aided Detection in Diagnostic Imaging subcommittee (CADSC) "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." This effort preceded the Radiological Devices Advisory Panel meeting in March of 2008, convened by the U.S. Food and Drug Administration (FDA), which discussed the current experience with and concerns about CAD devices and sought regulatory and scientific recommendations from outside experts on CAD. The AAPM CADSC has a membership and participants of diverse backgrounds and experiences (radiologists, CAD manufacturers, academic researchers, and government entities) (see the Appendix). The CADSC has conducted extensive discussions and formed subgroups to address issues related to CAD in four major areas:

- Methodologies for evaluation of standalone 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 system.

The purpose of this paper is to review the current status of the work done on the last two topics (i.e., CAD QA and training practices), share the rationale for establishing formal

QA and training procedures for CAD use in clinical sites, propose preliminary recommendations on CAD QA and training procedures, and discuss the off-label use issue, as well as the next steps moving forward to the clinical arena (i.e., potential mechanisms for an effective implementation of a CAD QA program at clinical sites and research and funding opportunities). The proposed recommendations are the opinions of the members of the AAPM CADSC, and intended to be work items for AAPM task groups that will be formed to address QA and user training issues on CAD in the future. The work items may serve as a framework for the discussion and eventual design of detailed QA and training procedures for physicists and users of CAD. It is hoped that, by recognizing the need for appropriate CAD QA and user training to ensure the effectiveness of CAD in clinical practice, the community will begin to work on developing and implementing some of the procedures. The potential benefits of improved quality and scientific rigor of CAD in clinical use may eventually outweigh the costs associated with QA requirements and regulations.

II. RATIONALE AND RECOMMENDED PROCEDURES FOR QA OF CAD AT CLINICAL SITES

The goal of this section is to recommend a set of general guidelines on CAD QA procedures. Note that the proposed set of guidelines as contained in this paper has been established through discussions among a number of experts in CAD who worked through a series of questions (see Sec. II.A) and reviewed various scenarios in order to understand why a CAD QA program may be needed, what CAD QA would require in terms of effort and cost, and how CAD QA procedures could be performed. Sections II.B and II.C present detailed procedures required to assure that a CAD system functions and performs according to vendor's specifications and to assure consistency of its performance over time, respectively. Note that some of the recommendations are considered by the subcommittee to be reasonably easy and practical and can be implemented immediately by the end users. Some of the recommendations are considered to be "best practice" approaches, which could potentially provide a more accurate measure of CAD performance. However, such "best practice" approaches may require significant effort and additional tools. The advantages and disadvantages of these "best practical approaches" are discussed in Secs. II.A–II.C and summarized in Sec. VI.

It should be noted that the testing procedures discussed in Sec. II do not cover all the testing that may be required for all possible CAD systems. For example, additional QA procedures may be required for CAD systems intended for determining the likelihood of malignancy for a suspicious abnormality, and estimating the risk of developing a disease or not responding to treatment. On the other hand, many medical imaging analytic software packages that are used to assist in diagnosis and decision making may not necessarily be labeled "CAD" (for example, software to segment lesions or superimpose a parametric color map on a displayed image). These packages could potentially benefit from QA testing. Although the specifics and extent of the testing may vary from one application to another and we often use the more commonly used lesion detection CAD as examples, the general principles and considerations discussed herein should be applicable to the various types of computer-assisted diagnostic software tools.

II.A. General guidelines on QA for clinical CAD systems

Seven questions were used to guide development and recommendation of a set of general guidelines for CAD QA procedures.

Q1. When does a QA procedure need to be performed?
Q2. Who should perform the test?
Q3. What are the tools and materials that may be used to perform the QA tests?
Q4. What are the performance measures to be used?
Q5. What QA results should be captured and how should they be reported?
Q6. What are the criteria of success/failure of the tests (or minimal requirements)?
Q7. What should be done if the CAD QA test fails?

Responses to these questions were not made specific to a particular CAD application and are detailed below.

Q1. When does a QA procedure need to be performed?

R1. CAD QA test is recommended for the following situations:

a. Performance verification/conformance to specifications (as stipulated by the manufacturers via the user manual, promotional materials, scientific publications, etc.) at the first installation or subsequent CAD software upgrades, including operating system upgrades,
b. Replacement or upgrade of imaging system hardware or software components that affect image quality (e.g., x-ray tube, energy source, detector, or image reconstruction algorithm),
c. Check for interaction or impact on the CAD software when other software or upgrades are installed on the same workstation on which the CAD system resides,
d. Any experience of performance deviation from the original specifications,
e. Specific procedures identified as necessary to be performed on a regular basis (e.g., monthly, quarterly, yearly) to assure performance consistency,
f. Change in image acquisition parameters, e.g., change in mAs, kVp, image reconstruction algorithm.

Q2. Who should perform the test?

R2. CAD QA should be performed by designated quality assurance personnel with adequate training to perform the QA procedures, e.g., technologist, medical physicist, radiologist, and manufacturer.

Q3. What are the tools and materials that may be used to perform the QA tests?

R3. CAD QA may use the following materials and tools:

a. Software tools provided and validated by CAD manufacturers or third party providers for:
   - Saving and retrieving user-selected image data to and from user-designated QA directories to ensure that the QA directories are separated from the local patient image storage or long-term patient archive of the clinical site,
   - Marking, validating, and storing appropriate reference standards (e.g., lesion location, pathology, and boundary) for the CAD system of a given application (e.g., detection, diagnosis, segmentation) in a CAD manufacturer-specified format for automated scoring of the CAD output (scoring refers to, e.g., the process of determining whether a CAD mark is a true positive or false positive, or the accuracy of a quantitative measure, by comparison with the truth or reference standard based on some defined measures or rules),
   - Storing CAD system output, such as CAD mark locations, lesion malignancy estimates, etc., if it is not an option available to users already,
   - Reading CAD system output, which may be stored as Digital Imaging and Communication in Medicine (DICOM) standard objects/files, for any user-selected cases (such a reading tool is important for many purposes, especially QA, and is not commonly available like a DICOM image reader),
   - Displaying CAD results for QA purposes,
   - Calculating performance measures (see responses/recommendations to Q4).
b. A set of clinical images (of appropriate version, e.g., “for processing” or “for presentation,” depending on the required input of the CAD system) with appropriate known reference standards provided by the CAD manufacturer or collected at the clinical site for the CAD system as described above.
c. Phantoms that are validated for a specific CAD QA purpose with a validated test procedure.

Q4. What are the performance measures to be used?

R4. Performance measures depend on the application of the CAD system and the purpose of the QA tests. For testing CAD system performance, the measures may include, for example:

a. CAD detection mark rate over time and stored in the CAD system (display time-series graph at the user demand for QA),
b. Sensitivity, i.e., the ratio of correctly marked lesions to the total number of lesions evaluated, relative to the reference standard. In some applications, sensitivity needs to be calculated for individual lesion types (e.g., mass, microcalcifications) or within certain regions (e.g., lung nodules behind the heart on chest radiographs, in the hilar region versus lung periphery for lung CT),
c. Segmentation accuracy, using an appropriate measure for the application (e.g., comparison of size or volume with the reference standard, an appropriate overlap measure or boundary comparison with the reference standard),
d. Malignancy/benignity assessment accuracy [e.g., area under the ROC curve or (sensitivity, specificity) pair].

The following should be noted:

(1) When assessing detection sensitivity, it is advisable to report the characteristics of the cancers evaluated, for example, in terms of the number of cases for different cancer types, sizes of the lesions, and other characteristics, which may affect the sensitivity.
(2) Other applications such as characterization and segmentation may also be affected by lesion characteristics (e.g., lesion size, spiculated lesions vs nonspiculated lesions, etc.). Descriptors of the data sets should be part of the stored reference standard.
(3) QA tests described here are mostly to track relative performance changes within the same data set. End users should be cautioned that the QA performance results may not be compared directly to performances estimated for the “patient population” reported by the CAD manufacturers or by other studies, unless the data set is representative of a similar population and the sample size is sufficiently large. The user should also be cautioned that sensitivity and specificity performance will change with each version of software as documented in the labeling for each version. CAD output such as the locations of CAD marks or segmentation boundaries may also change.

Q5. What QA results should be captured and how should they be reported?

R5. The following data may be captured, tracked, and reported from either visual assessment or automated computer analysis:

a. Log files with:
   - Date and reasons to perform the QA tests,
   - Personnel who perform the tests,
   - QA procedures or tasks performed,
   - Locations where the test data and results are stored if electronic storage is employed.
b. Performance measures (see responses to Q4)
c. Screen shots capturing CAD results on the image
d. Noticeable changes in performance over time or when compared with product specifications
e. Failure in communication between CAD systems and other necessary devices (e.g., acquisition system, display workstation, other imaging devices)
f. Record of intermediate results that are deemed to be necessary or critical (e.g., segmentation of the organ, feature values).

Q6. What are the criteria of success/failure of the tests (or minimal requirements)?

R6. The tolerance limits or criteria of success/failure for each specific QA test will depend on the application, sample size of the test data set, and require further research and discussion. As general principles:

a. For periodic QA procedures, hardware or software changes, or experience of performance deviations: definition of the level of change in performance (if any) as acceptable, passable, worrisome, or unacceptable,
b. For CAD system upgrades: tolerance limits if performance decreases but no limit on improvement.

Q7. What should be done if the CAD QA test fails?

R7. If the CAD QA procedure fails, the following actions should be considered:

a. Trouble-shooting of whether a change in hardware or software, that may or may not be part of the CAD system (e.g., a display workstation software upgrade or a change in image detector), affects CAD performance,
b. If the failure is a result of a discrepancy between the site's own test data and other data, investigation of whether the site's own test data set is representative of the larger patient population,
c. Investigation to assess impact of changes in performance and alert site users of potential problems,
d. Contact device developer or manufacturer.

II.B. Assure functionality and performance of CAD according to vendor's specifications

II.B.1. Acceptance testing at installation and subsequent upgrades of the CAD system

Currently, CAD QA procedures (or acceptance testing) are performed by most CAD manufacturers at the initial installation and subsequent CAD software upgrades based on the manufacturers' own procedures and using the manufacturers' test sets. These data sets are usually small. Acceptance testing procedures ensure that the CAD or upgraded CAD is installed properly, functions correctly, and integrates properly with image acquisition and display hardware and software (e.g., DICOM compatibility). Such procedures should be adopted by all CAD manufacturers including those that market and deliver the CAD system through the Internet. The test results should be documented as a reference for future QA of consistency. There are also situations where the clinical sites want to repeat such acceptance testing procedures using either the manufacturer's sample test set or their own independent data set(s) (see requirements on data sets in Sec. II.B.3). It is therefore important that all clinical sites be trained and encouraged to perform CAD acceptance testing. Such QA procedures can be implemented immediately at clinical sites.

Currently, results of such acceptance testing or QA procedures are not documented at clinical sites. It is recommended that the test records should be kept (for example, in the form of a log book or archived in a designated QA database) to allow subsequent review and comparison. As recommended in general guidelines R5 (see Sec. II.A), these may include the CAD final output results as well as intermediate results, pertinent notes on the results obtained, and any relevant information on problems encountered during the installation or upgrades of the CAD system. QA personnel at the clinical sites should independently evaluate performance in accordance with the specifications given by the CAD manufacturer. Adequate training and written instructions provided by the manufacturer are necessary to ensure that site personnel can perform QA tests correctly.

For an upgrade of the CAD software, the manufacturer should inform the site personnel of any expected changes. The site personnel should be advised to compare the CAD results before and after the upgrade and communicate any unexpected differences. It is important to understand whether or not the differences are intended by the upgrade. However, if the test set is small, one may not see the expected differences (e.g., performance improvement) due to statistical uncertainties. Longer-term tracking such as the CAD mark rate (described below as a routine automatic QA procedure) and the end users' qualitative or quantitative reporting of any performance change (e.g., radiologists' observed sensitivity of the CAD system) may be needed.

II.B.2. Installation of non-CAD software on the same workstation

For some modalities such as mammography, only the software for display of the CAD output is provided on the workstation. However, if a clinical workstation is a part of the picture archiving and communication system (PACS), the workstation may also offer software for display or analysis of other image modalities. The different software may have unforeseen effects on one another, most likely resulting from changes in the system configurations when software is installed or upgraded. Therefore, when other software (either CAD or non-CAD) is installed on a workstation or PACS system, it is important to perform QA procedures of all software (including CAD) that resides on the workstation in order to verify that it still functions correctly and in accordance with performance specifications. The results of QA tests for CAD should be identical before and after installation of other software on the workstation. As such, comparison of previous and current CAD marks, or quantitative outputs for some CAD systems in a few test images, should suffice. Such QA procedures can be implemented immediately and encouraged at clinical sites.

II.B.3. CAD QA test sets

A fixed set of test cases is recommended for acceptance testing at the first installation and subsequent upgrades of the CAD software. For convenient QA testing of consistency in performance, use of a CAD manufacturer-provided data set is appropriate. This data set should represent the typical performance of the system, and the specific CAD behavior on those cases, including truth, locations of CAD marks, and resulting sensitivity and specificity (or other quantitative output) should be well documented. Such test sets can be made available immediately to the end users.

Because a manufacturer's sample is likely to be limited and thus cannot be considered representative of the patient population seen at individual clinical sites, clinical sites are also encouraged to collect their own sets of test cases for performance assessment. It is advisable to collect a sufficient number of cases and use the automatic analysis tools to obtain a reliable estimate of the average performance. The data sets may come from existing databases at that site or third party sites, or they may be collected prospectively after installation of the CAD system. For a user-collected independent test set, the user is responsible for collecting reference standards required to assess the CAD system performance, following instructions and using tools provided by the CAD manufacturer (e.g., a bounding box or an outline of the lesion for marking lesion location stored in a specific data format and directories). The reference standard for each case has to be established by a clinician or other qualified personnel.

It is recognized that collecting such independent test data sets will require both additional effort and experience. The motivation for this additional, independent testing is that it will not only provide a higher level of QA to monitor CAD system performance in the local patient population, but it may also improve the effectiveness of end users' use of CAD by increasing their understanding of CAD performance in the local population.

II.B.4. Manual vs automatic assessment of CAD QA performance

The CADSC group members recognize the importance of both visual and automatic evaluation of CAD performance. Visual assessment of individual cases allows users to review images along with CAD output. It provides users the opportunity to assess performance specific to lesion type and identify any quality issues related to the display of CAD results. For lesion detection CAD systems, for example, users can track the performance by counting and visually assessing the CAD marks. However, it will be impractical, if not impossible, to score manually the overall performance of the CAD system in a substantially large data set or compare the results quantitatively or statistically with some expected performance level. Therefore, it is recommended that CAD system manufacturers provide built-in automatic scoring functionality. This automatic QA will run the test cases in batch mode and report the resulting sensitivity and specificity relative to the collected reference standards. On the other hand, even with fully automatic QA, it is recommended that each clinical site performs visual checks to ensure that the output of the CAD system, such as lesion segmentation or CAD mark locations, is reasonable and consistent. Automatic analysis would require proper evaluation software tools and validation of the reference standards used for scoring (see details in Sec. II.A responses and recommendations to Q3). Periodic manual/visual checks can be advocated and implemented immediately at clinical sites. Automatic assessment of CAD QA performance would be the best practice approach but may require significant effort from CAD manufacturers or any pertinent parties to implement or incorporate the functionality as QA tools.

It is recognized that even with fully automatic scoring and reporting, the interpretation of test results for locally collected data may be more complex than testing with the manufacturer's data set. For example, the local CAD performance may differ when compared against the manufacturer's claims or tests conducted at other sites. The difference may be attributed to many factors, such as (a) a true difference in the CAD performance if the test set is sufficiently large and representative of the local population, (b) difference in the image quality between the local data set and the manufacturer's data set, (c) statistical uncertainty due to the sample size of the independent test set, or (d) one or both of the test sets are biased. As discussed elsewhere in this report, such variability may or may not necessarily indicate problems with the CAD system, and the interpretation of such test results in the absolute sense requires caution. Some guidelines on independent testing, sample size, and quality of the test data set, and statistical interpretation of test results may be found in a report to be published by the AAPM CADSC Groups 1 and 2[3] and in the literature.[4,5]

II.C. Assure consistency of CAD system performance over time

Tracking the performance of a CAD system over time is an important QA procedure to ensure that the CAD system is consistent within its performance specifications. Tracking the performance after initial installation could help monitor, for example, whether variability that may occur in a component of the imaging chain affects the performance of the CAD system, or whether an upgrade by the CAD manufacturer actually improves the performance in the local patient population. CAD systems for lesion detection are discussed below as an example.

II.C.1. Tracking the number of CAD marks per image over time

Tracking the number of CAD marks per image over time on routine clinical cases is the most efficient and effective way to detect any unusual behavior of a CAD detection system. The QA procedure to track the running average number of CAD marks per image over N consecutive patients at any given time can be completely automated. The CAD mark rate monitor (in the form of a graph giving the average number of CAD marks as a function of time) may be inspected periodically by QA personnel. Visual inspection of the time-series plot could greatly inform the end users on the stability of the CAD system over time. An unexpected trend or sudden increase/decrease in the CAD mark rate will be a cause of concern and warrant investigation. If the QA test on an archived data set does not show a change, but deleterious changes are observed with routine clinical cases, the imaging chain should be evaluated for changes in image acquisition and processing. Such a QA procedure is reasonably easy and practical to implement, but it requires development of an automated recording and storage capability for the number of CAD marks for each image as well as software to generate and display a graph of the QA results.

II.C.2. Checking sensitivity of CAD system

The sensitivity of a CAD system is another important CAD QA index. As described in Secs. II.B.3 and II.B.4, it is important to evaluate the CAD performance on a clinically representative database. However, additional effort is required to establish an independent database of abnormal cases and to perform automatic performance evaluation. It should be noted that once the database is established, the database can be used to check the sensitivity of the CAD system both for new installation and after subsequent upgrades. An alternative to accumulating consecutive abnormal cases by the end users is to use a third-party or public database, if available.

In addition, CAD manufacturers could provide software tools that allow users to collect the data set and reference truths and calculate the sensitivity as described in the general guidelines (see Sec. II.A) to facilitate such evaluation in case the end users want to evaluate the sensitivity of the CAD system on their own patient population. Whether the sensitivity assessment on a nonmanufacturer-provided data set should be included as an acceptance or a routine QA test will be a topic of discussion for the CAD community.

III. RATIONALE AND RECOMMENDED PROCEDURES FOR TRAINING RADIOLOGISTS AND TECHNOLOGISTS ON CAD USE

To attain the full benefit of CAD, CAD users should be trained in the use of CAD. However, best practices have not been established for CAD training. Radiologists have been anecdotally somewhat resistant to formal CAD training. The purpose of the training is to educate the physicians on the intended use of the CAD system, although it is recognized that much of the physician's education occurs during clinical use. For example, research shows that the learning curve for radiologist use of CAD changes over the course of a year. In one study, breast radiologists initially increased their recall rate when using CAD, but over a year, the recall rate decreased to near the level before CAD implementation.[6]

Understanding what a CAD system has been designed to do and its limitations is important because using CAD beyond its intended capability can lead to ineffective or even harmful results, so user training is necessary. Training may be costly, and changes in personnel often mean that more training is necessary. Possible solutions include web-based training, but there is no substitute for one-on-one training and case-based training, which may be incorporated as a part of residency training. In recent years, CAD has been added to RSNA training modules for mammography.

Some of the important questions relating to training of CAD use that could be addressed through research include:

- How can we encourage radiologists to spend time getting trained?
- Does the absence of training impair CAD performance in the clinic?
- Is there an association between radiologist performance and attitudes toward CAD before and after training?

III.A. Importance of understanding the effect of improper use of CAD on sensitivity and specificity

The need of user training can be exemplified by CAD systems for lesion detection. It is critical for end users to understand the sensitivity and the average number of false positives per image in order to work effectively with these CAD systems. Each CAD system will have its own unique strengths and weaknesses depending on how it is developed and the data are used as a part of the development process. For example, the characteristics of the true-positive and false-positive detections may vary significantly from one system to another, even for CAD systems that have a similar performance in overall sensitivity and specificity. Proper training with clinical examples may significantly reduce the difference in learning curves among end users and potentially shorten the learning time.

Understanding the specificity and the types of false-positive detections is important for end users to use a CAD system efficiently and effectively, especially when false-positive detections are different from the typical false positives detected by clinicians. Proper training can help end users learn to dismiss certain types of false-positive detections, avoiding unnecessary workup or even biopsy.

The use of CAD systems in unapproved reading modes (e.g., off-label use) can adversely affect sensitivity and specificity and is discussed in Sec. IV.

III.B. Frequency and type of training

Training at initial installation is recommended. Whether retraining will be needed when CAD is modified or improved may depend on the magnitude of the change. It would be helpful to have recommendations as to whether additional training is necessary. From the perspective of the clinical site, the question is whether the change is substantial. The vendor should provide detailed documentation regarding what has been changed and what the user may expect. The site should perform CAD acceptance testing (see Sec. II.B.1) with the same test set that was used for initial acceptance testing to verify whether the observed change in performance agrees with the vendor's description. The radiologists can inspect the CAD marks before and after the change on the onsite test set to learn how the CAD marks change (see Sec. II.B.4). Seeing the specific changes in the CAD marks on a number of test cases will be the best training for radiologists.

The American College of Radiology (ACR) is actively engaged in training radiologists. ACR has included CAD in their Breast Imaging Residency Training Curriculum, which may pave the way for CAD training in other applications when they become more prevalent. Radiologists will likely accept structured training and prefer the integration of CAD in existing clinical training courses, continuing medical education (CME) and accreditation, rather than being offered as a separate course. Variations in the performance of CAD systems should be considered in the design of training courses. Mammography Quality Standards Act (MQSA) or other regulatory processes may impose training requirements for mammography and other CAD, respectively.

Technologist training may be useful, especially if radiation dose or patient positioning affect CAD, which may be especially important for less mature CAD applications. CAD should be assessed per modality and application to determine whether technologist training is essential, an assessment that could be made by the American Society of Radiologic Technologists (ASRT).

There are no guidelines or research on the conduct of CAD training or how much training is needed. A few studies observed that training affected radiologists' performance with CAD.[6,7] However, more research is needed to investigate how much training is required and how much variation there is among radiologists of different clinical experiences or practicing in different settings. It is also unknown whether CME would be helpful and, if so, to what extent. This situation presents an opportunity for collaboration among AAPM, ACR, and RSNA to encourage training on CAD use, determine how much training is necessary, and organize appropriate refresher courses for the various purposes.

III.C. Training the vendors: Feedback on CAD performance

Clinician or end user feedback to the vendor is an interesting topic. In a sense this is mutual training, not only for the clinician but also for the vendor. For example, the radiologist could send the CAD vendor examples of frequently missed cases that could point out a weakness in the system. The CAD vendor will usually first ask whether the missed lesions are mostly malignant. If benign, and such lesions are not part of the intended use of the system, then this is an opportunity for the vendor to educate the radiologist that the system is not meant to detect benign findings. Continuous feedback from the users to the vendors (and vice versa) is crucial and should be encouraged since it can accelerate improvements in the CAD systems. This is an example where auditing (monitoring the CAD performance) at the users' sites will be useful for the vendors. If the users accumulate statistics on how the CAD system performs and record cases that are missed by CAD, the vendors may get feedback from these results. Such a process is probably partially being done already at some institutions as a part of an educational process, internal auditing, or research, so it may not cause a lot of additional work. However, the vendors have to provide functions to facilitate such a process, for example, recording callbacks that have no CAD marks, CAD marks that cause additional callbacks, recalling CAD marks on prior exams when a current exam is being read, etc. The last function will help radiologists identify if a newly diagnosed cancer was marked in the prior exam but ignored. This kind of feedback will be valuable as a continuing training process for the radiologists to learn how to trust or discard CAD marks.

III.D. Storing CAD marks long term for training/auditing purposes

It was suggested that AAPM recommend long-term storage of CAD results. In the not-too-distant future, when the electronic health record (EHR) is more widely implemented, there will be massive databases of images, results, and long-term follow-up. CAD output should be a part of the long-term electronic record to allow healthcare professionals, vendors, and researchers to perform a variety of interesting analyses. Radiologists have not wanted to store CAD results because of a concern over medicolegal issues, but this concern has mostly not been borne out in the courts.[8]

The storage of CAD marks has many potential advantages. It will facilitate automatic monitoring of the stability of the CAD system performance over time (see Sec. II.C.1). It may help the radiologists to learn the characteristics of dismissed CAD marks on prior exams that turn out to be true lesions in current exams. It is also possible that the CAD system can use previous readings to improve its current performance after being trained to do so.[9] Analysis of the CAD prospective performance in large populations will allow further research and development to improve the CAD systems. On the other hand, the current lack of audit trail of CAD may diminish CAD's role as a serious diagnostic tool, and even question its reliability and scientific vigor. The lack of records also makes it difficult to conduct large clinical trials with proper study designs to evaluate the impact of CAD in routine clinical practice.

Most mammography CAD actually analyze a "for processing," or raw image, that is not stored unless the clinical site chooses to store it. The radiologist reads a processed image known as "for presentation." The same may be true for digital chest examinations and lung CAD. Thus, once the raw image is deleted, it may not be possible to recreate prior CAD marks from scratch. The same paradigm may apply to CT colonography. For example, there may be a role for CT colonography CAD to use thinner slices than those used "for presentation." In such situations, it will be even more important to store the CAD marks as a part of the electronic record. The same considerations should be given to other CAD systems that provide quantitative image analysis results or decision support to clinicians, similar to what is being done for many other non-CAD diagnostic tools.

IV. OFF-LABEL USE

Off-label use of CAD is using CAD in a manner that is not specifically stated in the FDA-approved indications for the CAD system. Physicians may use any FDA-approved product off-label according to their professional judgment concerning the needs of their patients. However, there is a concern that CAD will be used off-label in ways mainly to improve physician productivity such that the diagnostic performance of the physician plus CAD could fall appreciably. The issue here is how to handle off-label use of CAD.

Off-label use is more of a problem for radiologists when CAD is less sensitive than the human reader. For example, if the CAD system is labeled to be used only in the second reader mode, using it in the first or concurrent reader mode could potentially result in a degradation in sensitivity. This would be of particular concern when the CAD marks on the image are a mixture of output from two systems, one with presumably higher sensitivity (e.g., microcalcification detection) and the other with lower sensitivity (e.g., mass detection) than the radiologists, as in mammography. This higher sensitivity of CAD microcalcification detection might mislead radiologists to be less vigilant with their own reading and inappropriately use CAD in the first or concurrent reader mode. While the off-label use of the CAD marks might not cause much harm for the detection of ductal carcinoma *in situ* (mainly microcalcifications), the sensitivity for detection of invasive cancer (mainly masses) might suffer.[10] The concurrent reader paradigm might lead to satisfaction-of-search error. For example, while focusing on the primary disease-of-interest such as a lung nodule or colonic polyp marked by the CAD system, radiologists might be distracted from searching for lesions that the CAD system misses and might overlook secondary findings on other parts of the image.

As a general principle, the CADSC group notes that off-label use of CAD is not a good practice. This recommendation should be stated unequivocally because the impact of CAD depends strongly on reading paradigm.

IV.A. Means and practicality of controlling the reading environment

Presently, the reading environment including the off-label use of CAD is not monitored or recorded. There are concerns that readers may convert second-read-approved CAD to concurrent read and concurrent-read-approved CAD to first read. It is difficult to know how to discourage such behavior. Training is one way to do so. Another is to record or track reading behavior. The readers will know that they are being monitored and will be less likely to cut corners and use CAD off-label. The recording could consist of storing reading times and clicks as an audit trail.

Another more intensive approach is to record the CAD reader behavior while interpreting the images without and with the use of CAD in more detail. For example, the radiologist's findings could be recorded prior to displaying the CAD output. That way the pre-CAD readings would be available for auditing. The CAD vendors could control the workflow by modifying the display protocols. However, there is a concern that such auditing by the workstations might not be practical because in some review workstations CAD does not control the reading environment. The display protocols and the ability to audit CAD usage are functions of the workstations, not the CAD systems. In such situations, collaboration between the CAD vendors and the workstation vendors, if they are not the same, will be crucial for implementing CAD in the workstation with the proper control of CAD output display. Some thought should be given to adding these requirements to Integrated Healthcare Enterprise (IHE) protocols. Another possibility is to record the CAD metadata either in the DICOM header or in an alternative structure such as Annotation Imaging Markup (AIM), part of the National Cancer Institute's caBIG® (cancer Biomedical Informatics Grid®) initiative.[11]

The situation is further complicated because the physician may use multiple software tools when interpreting a case. For example, the radiologist often prepares final study reports using a dictation system; there may be no physical connection between the CAD and the dictation software, which raises the question as to whether the reporting functions of specialty CAD workstations are wasted.

Another issue is who would perform the audit. Manufacturers have no authority to perform audits. If the FDA or another entity were given the authority (similar to MQSA), they could audit reader performance with CAD. To have a more practical solution to off-label use, more feedback from physicians about how to accomplish this goal in a positive, supportive manner will be needed. How to design workflow to prevent off-label use of CAD is a good research topic.

IV.B. Standardization

Standardization can be helpful for advancing the clinical implementation of new imaging technologies and third party QA software tools, and reduce user training efforts for both QA personnel and clinical users. Examples of standardizations for CAD include technical and clinical functions such as file formats, reported data elements, and application programming interfaces (APIs) for PACS integration of CAD. Standardization, however, can be undesirable for the vendor community because it can become more difficult to differentiate one's product from that of others. A possible solution is to minimize the subset of functions on the workstation that requires standardization. Standardization is a driver to conformity and hence must be made attractive to the vendor. To identify the minimal requirements for CAD standardization, a community of users and vendors should form a forum in which all interested parties in the CAD community (manufacturers, healthcare professionals, academic researchers, government entities, public advocate organizations, and regulators) can communicate efficiently. It is suggested that the Medical Imaging and Technology Alliance (MITA) organizes and participates in such a forum.

Standardization requires extensive discussions to achieve acceptance or "buy-in." A successful example in this area is the development of mammography display and printing quality control standards. The FDA asked the vendor community for consistent quality control recommendations, so the vendors prepared such recommendations through MITA. There are also practical limitations to incorporating requirements, and an underlying set of standards would help. For example, the IHE (http://www.ihe.net/) helps narrow the DICOM standard to make it more feasible for vendors to implement. The discussion forum of the CAD community can work together to identify relevant standards.

Some of the recommendations regarding standardization for CAD are as follows: (a) CAD output should be stored for quality assessment and research. The CAD output should be stored as metadata either within the DICOM image header or using another standardized system such as caBIG's AIM data structure. (b) CAD should be made a part of hanging protocols in accordance with indications for use of the CAD to enable internal auditing if necessary. (c) Standardization of certain aspects of CAD implementation and usage may be helpful. Survey and discussion should be conducted to determine what standards exist that might be relevant for CAD and what other standards may be needed. The goal is to rely on existing standards to integrate QA into CAD and limit the number of new standards that would need to be developed.

V. FUTURE CAD QA PROCEDURES: POTENTIAL APPLICATIONS AND RESEARCH OPPORTUNITIES

There are opportunities for QA research all along the CAD development and utilization pipeline. Research with phantoms and simulated images may reduce the burden of human data collection for certain QA procedures and CAD development. As CAD moves from development to deployment, there are research opportunities for QA in the healthcare setting and in the evaluation of training and performance of clinicians using CAD. After CAD has been established in a medical practice, utilization of large electronic health records will facilitate monitoring CAD performance changes over time and evaluating effectiveness of CAD.

A majority of the topics covered in Secs. II–IV require further research, for instance, to refine the design and mechanism of the recommended procedures for implementation of CAD QA and training, to define specific or alternative procedures, to expand QA to all types of CAD applications, to develop standardized QA procedures and tolerance limits for each type of CAD system, and to understand and evaluate the cost and efficacy (e.g., impact on workflow and clinical outcomes) of QA procedures. Such research efforts will certainly require collaboration from all parties involved in CAD (i.e., device manufacturers, clinicians, academic researchers, government entities) as well as funding.

This section covers other potential research opportunities, for example, alternative procedures using phantom and simulated images for CAD QA, the assessment of radiologist performance and training methods, and the use of national EHR.

V.A. Phantoms and simulated images for CAD QA

Physical phantoms, partially digital phantoms (e.g., simulated lesions digitally superimposed on real patient images), or totally digital phantoms (e.g., digitally simulated images and lesions) are increasingly used for device design and development and for understanding limitations of certain devices or procedures. It may be envisioned that such phantoms could also be employed for certain types of CAD QA procedures instead of collecting a large set of real patient images with reference truth. For example, phantom images may be useful for monitoring CAD software consistency or reproducibility over time. Of course, the practicality of a phantom approach depends on many factors. One important question is whether the CAD system response to the phantom images is close, at least in a relative sense, to its response to real patient data. The use of phantoms will most likely lower the burden or cost to the end users and/or CAD manufacturers, which makes it an attractive alternative for consideration. To date, very limited research has been performed on the use of physical or partially/totally digital phantom image databases for QA procedures of CAD products. This may be explained by the very limited number of available phantoms and, to our knowledge, none have been developed and accepted for the purpose of testing CAD systems.

Physical or partially/totally digital phantom image databases are worthy of consideration as potential resources for CAD QA procedures. Once a simulated lesion database is developed, it can be reused at different sites and can substantially reduce the effort for clinical image collection. Physical phantoms can capture the effect of variations in the imaging chain on CAD performance, but the limitation of physical phantoms is that it is cost-inhibitive to build a large set of different physical phantoms that will realistically cover the variations in the anatomical structures and lesion characteristics in the patient population. If the purpose of using the phantom images is limited to certain QA procedures that monitor relative changes, the requirements on the sample size and their similarity to real patient characteristics may be less stringent.

The CADSC group recognizes that the design and validation of physical or digital phantoms will require long-term research and development efforts that include consideration of many issues such as:

- Designing phantoms of a specific anatomical region by incorporating statistical variations in phantom structures representative of patient population for a given CAD application,
- Designing lesion features representative of patient population for a given CAD application,
- Defining relevant QA tests and the sample size needed for the tests,
- Developing QA procedure protocols,
- Validating QA test protocols using phantom(s),
- Implementing a QA mode on CAD systems by manufacturers,
- Defining pass/fail criteria for QA tests using phantoms.

The purpose of the discussion is to stimulate interest in the research and development of advanced phantoms or lesion models for specific CAD applications and the validation of phantom use in certain associated QA procedure protocols. If the phantom approach is developed and accepted by the CAD community, it may facilitate implementation of QA procedures.

V.B. Assessment of radiologist performance and training methods

The rationale and procedures for training radiologists were discussed in Sec. III. This section describes two complementary training processes.

First, at the introduction of a new CAD system in each clinic, the radiologists may be presented with a sufficiently large set of cases with known lesions that capture a range of difficulty without CAD and with CAD. Study of the CAD output on real images may familiarize the radiologist with the process and provide certain understanding of the CAD output. Questions associated with this training process include: (a) how many cases would be necessary so that the training is effective, (b) should the cases be provided by the CAD vendor or selected from the patient population as seen in the users' respective practices? The former may be efficient, while the latter may have the advantage that the CAD marks will reflect the specific image quality and characteristics in the local clinic. Another question is whether retraining is needed at each upgrade of the CAD system.

Second, during clinical use of a CAD detection system, for example, the radiologists should periodically review workups and biopsies that result from CAD marks and may display CAD marks on prior exams in their routine reading. The review may help reduce future overcalls due to false-positive CAD marks and reinforce their confidence in true-positive CAD marks. The display of previous CAD marks on the prior exams may help the radiologists learn the properties of early subtle lesions that may have been marked by the CAD system and thus reduce incorrect dismissal of true-positive marks in the future. In addition, archiving the CAD mark locations of every examination, rather than deleting them as in current practice, could better document the reasons than a dictated note for recalling a patient and defining a specific focal lesion for the additional workup. Research on the potential benefits of these procedures would help guide clinical practice.

To determine the proper mechanism for training the radiologists on the use of CAD, research investigations will be needed to address the questions above. It will also be interesting to determine the time period or number of cases that is necessary for a radiologist (having received training or not) to become consistent and most effective (i.e., optimal performance) with CAD. It is noted that the feedback of results of CAD in clinical use will require long-term storage and easy recall of CAD marks, as discussed in Secs. II–IV.

Training of radiologists and technologists may also be required to optimize the preparation of patients prior to being imaged by systems that utilize CAD. For example, in CT colonography, patients need to have an appropriate bowel preparation and at the time of scanning need to have adequate colonic distension and image resolution. Patient preparation and acquisition-specific issues that bear upon the performance of CAD should also be investigated.

In addition to training, other methods that may improve radiologists' proper use of CAD marks should be considered. For example, radiologists' interpretation may be improved with the assistance of content-based image retrieval or with probability estimates attached to the CAD marks. Although some work has been published in these areas,[12] much more research will be needed to explore the different approaches that may improve the effectiveness of radiologists using CAD systems.

V.C. Monitoring CAD performance changes over time and effectiveness using large electronic health records

It is envisioned that the use of EHR could be a means for government entities (e.g., MQSA for mammography), hospitals, and physician practices to monitor CAD performance (e.g., average number of marks per image) and its effect on radiologist performance (e.g., average number of CAD marks selected by the radiologist as abnormalities requiring additional workup). Electronic connection with other clinical information such as patient outcomes and long-term follow-up data across hospitals may facilitate assessment of the impact of CAD on healthcare. Recording the version of the CAD system as well as the imaging system model and image acquisition parameters would allow relevant stratified analyses (e.g., per type of scanner used, per geographic region, according to radiation exposure or patient preparation). Use of EHR may provide useful and precise information on the impact of CAD on the overall and subgroup performances and thus help guide further improvement of QA procedures and CAD use, as well as CAD research and development. FDA currently does not monitor CAD systems, and there are economic and medicolegal concerns considered as barriers that prevent recordkeeping of CAD in clinical practice. Involvement of governmental, professional, and patient groups could be a significant part of introducing QA to CAD.

V.D. Funding opportunities

Funding is a critical component for moving forward research and development efforts on QA of CAD. It would likely be challenging to obtain grants for such efforts because the topic may be perceived as lacking scientific innovation. However, the key is to emphasize the practical clinical significance for establishing QA of CAD. It is increasingly recognized that significant health benefits can accrue from quality improvement.[13,14] Some of the funding agencies focus more on basic research than on clinical implementation. Identification of alternative funding sources such as those emphasizing healthcare improvement and initiatives would be beneficial. Proof of the benefits of auditing is potentially a fruitful research focus.

VI. DISCUSSION AND SUMMARY

This paper discusses the rationale and recommendations for QA procedures and training clinicians and clinical site personnel regarding CAD systems used in clinical practice. The QA procedures encompass principles for QA testing at installation or upgrade of CAD and procedures to track the consistency of CAD performance over time. Some of the procedures are considered as practical and can be implemented immediately and some are considered as "best practice." These "best practice" approaches may require studies to assess their effectiveness before implementation. For example, the paper outlines the important roles of both manual (visual review) and automatic analysis of CAD output in QA processes. The visual review approach can be implemented immediately. However, the automated approach cannot be implemented immediately because it requires software tools, as discussed in Sec. II.A. In addition, the automatic analysis is recommended to be performed on a sufficiently large data set in order to obtain meaningful outcomes. This may require clinical sites to collect a large number of cancer or disease cases for estimating CAD sensitivity. Although the analysis can be performed automatically using the software provided by the manufacturers or a third party, the case collection process is rather manual and labor intense. For sites with a smaller patient volume and limited resources, it will take a relatively long time to accumulate enough cancer cases. The CADSC group recognizes that this process may not be practical for routine QA in all clinical sites. However, it is worthwhile to conduct further research to better understand its cost-benefit tradeoffs as a "best practice" approach, as well as to search for less burdensome methods that can facilitate the case collection, for example, studying the minimum criteria for establishing reference standards and the minimum number of cases required for the specific QA purpose. In addition, the CAD community may collaborate to establish a common data depository that includes geographic and demographic information to allow users to select a case mix to simulate the local patient population, thereby reducing the burden on individual sites. Case sets collected by a third party that meet quality requirements may also be evaluated as an option. It is expected that AAPM will form task groups in the future to provide specific guidelines on assessment of CAD systems, such as the criteria for reference standard, recommended number of cases, etc., to end users who are inexperienced in this area if this "best practice" approach is recommended for certain CAD QA purposes or for clinical sites that prefer to assess CAD performance in their local population. Independent of QA applications, such a database will be valuable for the purpose of user training. Adequate user training on CAD could substantially reduce the time to learn how to effectively work with CAD and potentially reduce the variation among the users in their performance over time.

The "best practice" approaches could potentially provide more accurate measure of the CAD performance at the local site and better training of the end users on CAD use. However, they cannot be implemented without additional effort or research. The potential impact on end-user clinical workflow efficiency should be considered when determining requirements and the level of training that is necessary for proper clinical use of CAD. Potential cost burdens on manufacturers have to be considered to garner vendor support for these activities and the final recommended QA procedures. It is envisioned that CAD manufacturers may play a major role in CAD QA by providing standardized software QA tools, which allow end users to perform QA tasks more efficiently and effectively. It may also be possible that a third party provider can provide the standardized tools more cost-effectively by reducing duplicate effort by individual CAD manufacturers. Either way, it is recognized that CAD manufacturers' participation is critical to the establishment of QA procedures and implementation of QA tools. It would be productive if the manufacturers share their experiences in QA, perhaps having MITA (http://www.nema.org/prod/med/) and the CADSC as a forum, and work jointly to design a common set of useful and practical QA procedures at installation and for periodic maintenance so that QA procedures will be performed in a consistent manner across products from different manufacturers. The goal is to bring CAD to an appropriate standard of quality and facilitate acceptance testing and monitoring the performance of CAD products at clinical sites.

QA of CAD systems for mammography may serve as a starting point. Study of the cost and benefit impact of a QA program for mammography CAD on patient care may provide a practical model to guide the design of QA programs for CAD in general. Participation of the MQSA should be proposed. Government entities' engagement in this effort, as well as development and implementation of regulations to include inspection of the QA program of CAD systems, could be an effective means to initiate QA monitoring of the proper use of CAD.

The CADSC recognizes that reimbursement drives much of the clinical use of CAD. While quality is important, in practice it may take a back seat to cost and reimbursement issues. However, we also believe that proactively seeking the highest level of quality in CAD and documenting its benefits in healthcare might convince payers to expand reimbursement for CAD. Quality measures could impact reimbursement for CAD, just as use of EHR and electronic prescribing affect reimbursement. It is expected that better QA could benefit the patients and the end users, and ultimately lead to wider adoption and reimbursement of CAD in clinical practice. These are important factors that will sustain and drive research and development in the CAD field.

Due to the diversity of CAD systems, the QA procedures discussed herein do not encompass all possible tests that could be important and necessary to perform on a CAD system. For example, a CAD system that provides not only prompting capabilities to some suspected lesions but also the outline of the findings that the system segments and determines as suspicious could be subjected to an additional layer of QA (i.e., those that will convey QA assessment of the segmentation accuracy). A CAD system that assists in estimating the likelihood of malignancy of a lesion, predicting the risk of a disease, or predicting the risk of recurrence or treatment response will require different validation and QA procedures. The recommendations are not intended to describe the specific procedures to be followed by the CAD end user or other designated QA personnel, but rather to provide guidelines or general approaches on how certain QA could take place and be performed effectively. It is certain that these guidelines will evolve through further discussions within the CAD community before a QA program can be implemented and through practical experiences after it is started clinically. Specific QA procedures for a given CAD application will have to be designed based on the application and the approved mode of use. Collaboration among the clinical and CAD community (i.e., academic researchers, healthcare professionals, software developers, manufacturers, and government entities) is essential for the initiation of CAD QA procedures and for their development, refinement, validation, and implementation in the clinical arena.

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 by the Intramural Research Program of the

National Institutes of Health (NIH), 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 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. S.G.A. receives research funding from Riverain Technologies through The University of Chicago. R.M.S. receives patent royalties and research support related to CAD from iCAD. M.T.F. and S.C.B.L. receive funding from Riverain Technologies through Georgetown University Medical Center.

APPENDIX: AAPM CAD SUBCOMMITTEE

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 and Harvard Medical School
Bin Zheng, University of Pittsburgh
Chuan Zhou, University of Michigan

[a] An opinion paper from the AAPM CAD Subcommittee.

[b] These two authors contributed equally to this work.

[c] Current address: 12300 Village Square Terrace, Rockville, Maryland 20852.

[d] Author to whom correspondence should be addressed. Electronic mail: chanhp@umich.edu; Telephone: 734-936-4357.

[1] 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).

[2] S. Astley, C. Quarterman, Y. Al Nuaimi, C. Chasser, I. Dukic, V. Hillier, F. Gilbert, U. Beetles, H. Deans, K. Duncan, M. Griffiths, G. Iyengar, A. Jain, M. Wilson, P. Griffiths, M. McGee, E. Singleton, S. Duffy, and C. Boggis, "Computer-aided detection in screening mammography: the impact of training on reader performance," in *Proceedings of the Seventh International Workshop on Digital Mammography, Durham, NC, 2004*, edited by E. D. Pisano (University of Carolina, Chapel Hill, NC, 2004), p. 231.

[3] N. Petrick, B. Sahiner, S. G. Armato III, A. Bert, L. Correale, S. Delsanto, M. T. Freedman, D. Fryd, D. Gur, L. Hadjiiski, Z. Huo, Y. Jiang, L. Morra, S. Paquerault, V. Raykar, M. Salganicoff, F. Samuelson, R. M. Summers, G. Tourassi, H. Yoshida, B. Zheng, C. Zhou, and H.-P. Chan, "Evaluation of computer-aided detection and diagnosis systems," Med. Phys. (2013) (accepted).

[4] B. Sahiner, "Methodologies for evaluation of standalone CAD system performance," Med. Phys. **39**, 3962 (2012) (abstract).

[5] N. Petrick, "Methodologies for evaluation of effects of CAD on users," Med. Phys. **39**, 3962 (2012) (abstract).

[6] 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).

[7] S. A. Taylor, D. Burling, M. Roddie, L. Honeyfield, J. McQuillan, P. Bassett, and S. Halligan, "Computer-aided detection for CT colonography: Incremental benefit of observer training," Br. J. Radiol. **81**, 180–186 (2008).

[8] R. J. Brenner, M. J. Ulissey, and R. M. Wilt, "Computer-aided detection as evidence in the courtroom: Potential implications of an appellate court's ruling," AJR, Am. J. Roentgenol. **186**, 48–51 (2006).

[9] S. Timp and N. Karssemeijer, "Interval change analysis to improve computer aided detection in mammography," Med. Image Anal. **10**, 82–95 (2006).

[10] J. J. Fenton, S. H. Taplin, P. A. Carney, L. Abraham, E. A. Sickles, C. D'Orsi, E. A. Berns, G. Cutter, R. E. Hendrick, W. E. Barlow, and J. G. Elmore, "Influence of computer-aided detection on performance of screening mammography," N. Engl. J. Med. **356**, 1399–1409 (2007).

[11] D. S. Channin, P. Mongkolwat, V. Kleper, K. Sepukar, and D. L. Rubin, "The caBIG annotation and image Markup project," J. Digit Imaging **23**, 217–225 (2010).

[12] B. Zheng, "Computer-aided diagnosis in mammography using content-based image retrieval approaches: Current status and future perspectives," Algorithms **2**, 828–849 (2009).

[13] K. K. Deshpande, R. M. Summers, R. L. Van Uitert, M. Franaszek, L. Brown, A. J. Dwyer, J. G. Fletcher, J. R. Choi, and P. J. Pickhardt, "Quality assessment for CT colonography: Validation of automated measurement of colonic distention and residual fluid," AJR, Am. J. Roentgenol. **189**, 1457–1463 (2007).

[14] R. L. Van Uitert, R. M. Summers, J. M. White, K. K. Deshpande, J. R. Choi, and P. J. Pickhardt, "Temporal and multiinstitutional quality assessment of CT colonography," AJR, Am. J. Roentgenol. **191**, 1503–1508 (2008).