I. Introduction ... 3
A. Purpose of the panel meeting ... 3
B. What do we mean by CAD? ... 4
C. How CAD devices for medical imaging work ... 7
D. Regulation of CAD devices ... 9
II. General Methodologies for Computer Aided Detection ... 12
A. Clinical use of computer aided detection ... 12
B. Nonclinical testing of CAD devices ... 13
C. Clinical testing of CAD devices ... 15
1. Hierarchical model of efficacy ... 15
2. Types of test and potential biases ... 16
3. Test dataset ... 18
4. Ground truth ... 20
5. Standalone performance testing ... 21
6. Reader performance testing ... 23
D. Primary concepts of statistical analysis ... 28
1. Study endpoints ... 28
2. ROC analysis ... 29
3. FROC analysis ... 32
E. Additional statistical issues ... 34
1. Comparator for reader studies ... 34
2. Sample size for retrospective reader studies ... 35
3. Statistical analysis of retrospective reader studies ... 35
4. Units of analysis ... 36
III. Mammography CAD devices ... 41
A. How mammography CAD devices are used ... 41
1. Screening for breast cancer ... 41
2. Patient characteristics in a screening populations ... 42
3. Mammography devices ... 43
4. Mammography interpretation ... 45
5. Performance measures of mammography ... 47
6. Radiologist errors on mammography examinations ... 48
7. Origin of mammography CAD devices ... 49
B. Mammography CAD devices approved by FDA ... 50
1. Currently approved devices ... 50
2. Data and information provided for original approval of mammography CAD devices ... 51
C. Published literature on approved devices ... 52
D. Clinical testing issues specific to mammography CAD ... 56
1. Ground truth ... 56
2. Standalone performance testing ... 56
3. Reader performance testing ... 57
4. Other issues for mammography CAD devices ... 60
IV. Lung CAD devices ... 61
A. Introduction of lung CAD into the clinical practice ... 61
1. Lung imaging ... 61

2. Lung cancer and pulmonary nodules 62
3. Chest x-ray and CT imaging of pulmonary nodules 63
4. CAD devices for chest x-rays and chest CT 65

B. How lung CAD devices have been approved 65
1. Chest x-ray CAD 65
2. CT lung CAD 66

C. Published literature 68
1. Chest x-ray CAD 68
2. CT lung CAD 68

D. Clinical testing issues specific to chest x-ray and CT lung CAD 69
1. Ground truth 69
2. Standalone performance testing 70
3. Reader performance testing 70

V. Colon CAD devices 74

A. Introduction of Colon CAD into the clinical practice 74
1. Background 74
2. Standard of care 74
3. CT Colonoscopy 75
4. Requirements for CTC interpretation 76
5. Origin of colon CAD devices 78

B. Cleared colon CAD devices 78

C. Published literature 79

D. Clinical testing issues specific to colon CAD 81
1. Ground truth 81
2. Standalone performance testing 82
3. Reader performance testing 83

VI. Other CAD Devices 87

A. Other CAD applications 87

B. When should devices be evaluated like CAD devices? 87

VII. Glossary 89

VIII. References for which full text is provided to Panel 92

IX. Endnotes (Bibliography) 97

Introduction

A. Purpose of the panel meeting

Over the last 10 years, FDA has seen an increasing number of submissions for computer-aided detection devices to be used with radiological images. As these devices have been developed and tested, the science related to these devices has grown, as has the clinical experience with such devices.

In this meeting, FDA seeks to obtain the advisory panel's recommendations on the data to be used to support the approval or clearance of CAD (computer-aided detection and/or diagnosis) software applied to radiological images, with emphasis on computer-aided detection devices. FDA is required by statute to be least burdensome in its requirements, so we ask the panel to consider what data are essential.

We start out by defining CAD (Section I). We provide the regulatory framework in which CAD devices must be shown to be safe and effective, i.e. there must be valid scientific evidence that when the CAD device is used in accordance with its intended use, there is reasonable assurance that the probable benefits will outweigh the probable risks, and reasonable assurance that there will be a clinically significant result in a significant portion of the target population.

Then we discuss the methods that are used to assess computer aided detection devices (Section II). We discuss how valid scientific evidence may come from standalone performance testing and/or reader performance testing using a test dataset that represents the target population appropriately. We proposed that the study design should match the intended use, and there should to be appropriate study endpoints, and ground truth. For reader studies, the reading paradigms should support the intended use.

Then we get into three major areas of discussion: mammography CAD, lung CAD, and colon CAD (Sections III, IV, and V). We provide background about each of these device types; then we ask the panel to discuss the types of testing that can be used for each device type, e.g. standalone and/or reader studies, test datasets, study endpoints, ground truth, and reader paradigm. Although there have been clinical protocols used for past approval of CAD devices, we ask for the panel's recommendations on how these devices should be evaluated in the future, in light of evolving scientific understanding and clinical experience.

Finally, we ask the panel to extend the thinking from mammography, lung, and colon CAD to other types of CAD devices, including diagnostic devices, that may exist now or in the future (Section VI).

B. What do we mean by CAD?

1. Background: clinical interpretation of radiological images

The performance and interpretation of diagnostic radiology tests is a complex process. The process typically begins with a *referral* (or request) from a non-radiologist physician to perform a specific radiology test. The test itself is typically *performed* by a technologist based on a protocol determined by a radiologist. Some tests are performed directly by the radiologist (e.g., breast ultrasound) or directly monitored by the radiologist (e.g., general ultrasound or MRI). Images, and associated image *data*, are *produced* using a variety of imaging modalities. The images are then *recorded* (i.e., stored) on some format such as film or digital archive. The radiologist then *interprets* the images.

The process of interpretation comprises four basic steps[1]: *detection*, *description*, *diagnosis* and *reporting*. Description can include quantitative measures such as finding size. The results are then *communicated* to the referring physician (typically in the form of a written/electronic report). Detection comprises identification of a "finding" on an image that is potentially abnormal and requires further scrutiny. The initial determination is whether or not the finding represents a normal structure. If not a normal structure, the next step is to describe the features of the finding. For example, if the finding on a mammogram is a mass, the following features should be described[2]: size; morphology (shape, margin and density); associated calcifications; associated findings, and location. The margin is further described as[3]: circumscribed; microlobulated; obscured; indistinct; or spiculated. Description obviously requires detection of individual features or sub-features. For example, the reader needs to detect spiculations in order to describe the margin of a mass as spiculated. Description also directly drives the diagnosis (that is why particular descriptive features have been developed over time). For example, if a breast mass has spiculated margins, the likelihood of malignancy is increased. Therefore, detection, description and diagnosis are intertwined in a complex manner. Once a finding has been detected and fully described, a diagnosis is then made. The diagnosis may comprise an "assessment category" (e.g. in mammography[4]). Similarly, assessment categories have been developed for CT Colonography.[5] In general, a diagnosis may be definitive (e.g., a finding is definitely benign or malignant) but is more likely in the form of a clinical action. A clinical action would include additional concurrent imaging, additional follow-up imaging, biopsy or even surgery. Following detection, description and diagnosis, a report of the findings is made and communicated to one or more referring clinicians and sometimes the patient him/herself.

2. Radiologist errors

Errors made by radiologists can be

- cognitive
- perceptual
- technical
- administrative.

Cognitive errors occur when a radiologist perceives an abnormality but misinterprets the nature or significance of the abnormality due to incomplete knowledge[6] or faulty reasoning or judgment.[7]

Perceptual errors have been defined to occur when a radiologist fails to see an abnormality at the time of interpretation and that abnormality is "evident," in retrospect, at a later time.[8] However, a more fundamental definition is that a radiologist makes a perceptual error when the radiologist (1) fails to perceive (2) material information (3) that is in fact "revealed" by the test. Many factors may contribute to the failure of a radiologist to perceive information that is in fact revealed by the radiology test. For example, a radiologist may interpret the test without adequate medical history, without any medical history or even with the wrong medical history. If the radiologist interprets a mammogram thinking the history is bloody nipple discharge, when in fact the history is a palpable mass in the upper outer quadrant of the right breast, the radiologist may fail to perceive a breast mass that is in fact visible in retrospect and would have been seen by the radiologist if he/she had had the correct history. When a chest CT is obtained with suboptimal technique, a nodule may be made less conspicuous, albeit still visible, and it may be overlooked. If the radiologist is extremely busy, and spends too little time examining the images, a visible abnormality may be overlooked. If the radiologist fails to compare the current examination with a prior examination or fails to review a recently performed study of the same body part with a different modality, this may cause him/her to overlook a visible abnormality. Numerous other possibilities exist. In some cases, more than one of these "contributing" factors may play a role in the failure of the radiologist to perceive an abnormality revealed by the test. In some cases, the radiologist does everything else properly—reviews all of the prior studies, has the complete patient history, the study is performed with perfect technique, etc., but a material finding revealed by the test is still overlooked. Such a case would constitute a "pure" error of perception. When any other factor(s) contribute(s) to a perceptual error, this is referred to as a "mixed" error of perception.

Technical errors occur when the quality of the examination is substandard due to an error in performance of the test, production of images, recording of images or the storage of images. There are numerous types of technical errors including incomplete imaging,[9] incorrect x-ray exposure,[10] poor patient positioning,[11] etc.

Administrative errors include imaging the wrong body part or wrong region of the body, losing the films or failing to store the digital images properly,[12] imaging the wrong patient, mislabeling the right side as the left side (or vice versa), failure to obtain or utilize adequate patient history or other clinical information, dictating a report with the wrong patient name, failure to compare the current test with prior tests, failure to correct an error in a dictated report, etc.

3. What is CAD?

The abbreviation "CAD" typically stands for computer-aided detection and/or computer-aided diagnosis. The term "computer-assisted" is used interchangeably with "computer-aided." Computer-aided diagnosis is also sometimes referred to as classification.[13] CAD devices can aid (or assist) with one or more parts of the image interpretation process for the purpose of reducing radiologist errors.

Many tasks in radiology are very difficult, such as finding small lesions among many layers of fibroglandular breast tissue. Other tasks can be very time consuming, such as reviewing a large number of CT slices. The intent of CAD devices may be to aid or assist practicing radiologists to perform these tasks more effectively, with greater safety, and/or faster.

Radiological CAD programs are often installed as part of a workstation platform, as an accessory to a radiological device such as magnetic resonance (MR) diagnostic device, Computed Tomography X-ray System (CT), or as an add-on operating function provided in a Picture Archiving and Communications System (PACS) or other system. CAD software will process the images collected or stored on such systems and save their output for subsequent display. Upon reviewing the digital images on a computer workstation (Figure 1[14]), clinicians have the option to display or query the CAD device output. The practicing clinicians determine whether the CAD devices are used and how they are used in their interpretations of images.

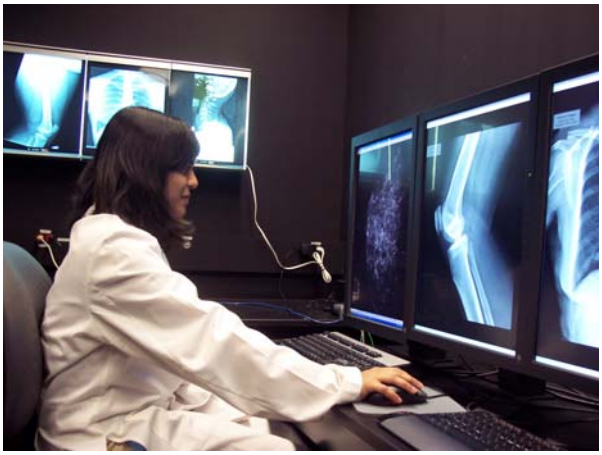


Figure 1: CAD devices may be installed on digital image display workstations

Devices that primarily influence the detection task are sometimes referred to as CADe (computer aided detection) devices, while those that primarily influence the diagnosis task are sometimes referred to as CADx (computer aided diagnostic) devices.[15] However, a CAD device may influence one or more of detection, description, diagnosis or reporting depending on the device design, testing, output, mode of interaction with the radiologist and intended use.

The prototypical CADx device is designed to process a specific finding (or region) of an image that has already been detected and determined to be (or contain) a potential abnormality, in order to characterize the finding. The detection may have been made by component CADe functionality and/or by the user. The device then characterizes the finding (e.g., a likelihood of malignancy score or a recommended clinical action such as biopsy or not biopsy, etc.) and/or describes the finding (e.g., describes various image features such as morphology), for the purpose of reducing cognitive errors. In sum, the prototypical CADx device tells the radiologist *what he/she is looking at*.

The prototypical "pure" CADe device is designed to process images to detect findings (or regions) on an image that may be (or contain) an abnormality. The term "abnormality" is here meant to include any imaging finding (or region) that is not a normal structure or normal variant. to CADx devices, however CADe devices automatically or internal (i.e., not available to the user) likelihood rating of est (ROI) in the image. ROIs deemed sufficiently abnormal are utputs of computer aided detection devices often appear as an al image (See Figure 2[16]). The device overlay marks, highlights, such a finding (or region) and brings it to the attention of the ist) for the purpose of reducing perceptual errors. Marks (or s "prompts" and may be of different shapes, colors, or pical CADe device tells the radiologist *where to look* on an

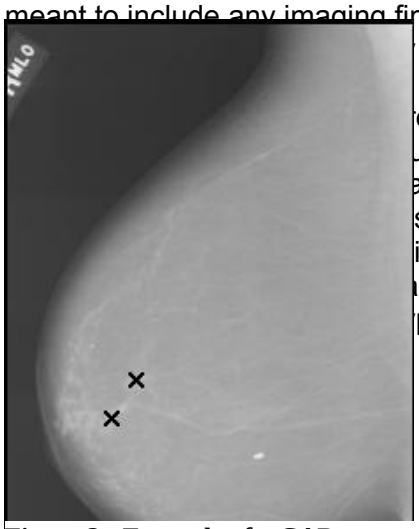


Figure 2: Example of a CAD marks which commonly appear as an overlay on a medical image.

It is important to keep in mind that the detection process will almost always include some component of diagnosis. A device designed to detect and mark both benign and malignant findings (of all sizes and types) would have little influence over the diagnosis task. On the other hand, a device designed to detect and mark only malignant findings may influence the diagnosis task.

CADe devices will mark some ROIs on images that correctly correspond to an abnormality (true positives). It will also mark other locations in error (false positives). Currently most CADe devices are set at a threshold that marks false positives at a rate far greater than the true positive rate. It is the task of the clinician to interpret which of these CADe marks identify an ROI that requires further scrutiny and which mark irrelevant ROIs containing only normal or benign structures.

C. How CAD devices for medical imaging work

Overview

A CAD device uses software algorithms to look for patterns in patient data or images that indicate a particular abnormality or imaging finding, such as a polyp. Just as a radiologist is trained to identify abnormalities using cases studies, the algorithms are first trained to identify patterns using data from a finite sample of patients who have that abnormality. This sample is called the training set. Once trained or optimized, the CAD device can then be used to indicate which data or ROIs from other new patients have patterns sufficiently similar to those in the training set. This will be described as detection (CADe). Additionally the CAD device can give a rating indicating how closely a specific part or ROI of the new patient data matched its trained pattern. This is computer aided diagnosis (CADx). The CAD device's pattern identification is intended to be sufficiently sensitive and specific to aid clinicians in their identification and/or assessment of abnormalities.

Note:

- CAD algorithms will tend to identify ROIs similar to ones in the training set. Using different sets of patient data to train the same CAD algorithm will yield CAD devices that match different patterns and patients.
- Different CAD algorithms will match different patterns and patients even if they are trained using the same set of patients.
- The patterns detected by a CAD are computed from complex transformations of the digital data or images and in general are not directly known or understood by the users of a CAD device.
- CAD algorithms typically are not adaptive; instead they are fixed, and they only change with new software revisions.

Specifics

Many medical imaging CAD devices internally perform the following steps when applied to a medical image, though there is substantial variation of how they are implemented.

1. *Image Processing*
   The image is enhanced, or processed to facilitate subsequent analysis by the particular CAD. This processing may include changes to contrast or brightness, histogram equalization, smoothing, or sharpening.

2. *Segmentation*
   Boundaries in the image are mapped to anatomy such as organs or other regions of interest (ROIs). Segmentation techniques differ among CAD devices and may involve techniques such as edge detection or region growing.

3. *Feature Calculation*
   The features of each segment or ROI in the image are calculated. Features are computer or human estimated quantities characterizing images, regions or pixels within an image. These features can include measures such as shape, size, volume, curvature, spatial frequency power, entropy, or fractal dimension. The number of such features can be in the hundreds. These features may or may not be calculable or understandable by the users. Different CAD devices will calculate and use different features, even for the same task.

4. *Classification/Discrimination*
   The features obtained in step 3 are fed to statistical learning algorithms, or classifiers. The classifiers will output a single value. Generally this value will indicate an estimated likelihood of disease of the region from which the features were calculated, but this will depend upon how the classifier was trained or optimized. Examples of classifiers include neural networks, decision trees, kernel machines, or simple linear classifiers. Different CAD devices use different classifiers. Some CAD devices may use multiple classifiers using different sets of features.

5. *Thresholding and Output*
   Generally a CADe device will place a mark on each segment or region of the image if the classifier output (estimated disease likelihood) for that region exceeds a certain threshold. A CADx device, when prompted on a particular ROI by a clinician, will present the output of the classifier for that ROI, either directly or in some scaled or quantized form.

Each of the above steps can be modified by including or eliminating methods or by changing numerical software parameters in each step. When CAD developers make such modifications they are developing, or “training” the CAD device. In general these changes are made such that the performance of the CAD device increases when used on a set of patients known as the training set. For example, the features that are calculated in step 3 may be chosen to maximize the sensitivity of a CAD device on a set of patients with breast cancer. This optimization or training can be performed using mathematical regression using a computer, performing an extensive search using a computer, having humans tweak parameters in the algorithm, or some combination thereof.

CAD devices can be "over-trained". If a CAD developer is not careful, the CAD device can become overly specific to the training data set. In this case the CAD device may perform nearly perfectly when used on the training set of patient images, but have very poor performance when used on any other set of patient images. There are automated training methods to avoid such problems, such as cross-validation. These methods require fully automated computer optimization of the CAD device.

Once CAD performance on the training data set is optimized, the CAD device can be applied to new sets of images. If the diseased statuses of a new set of patients are known, and these images have not been used in the optimization of the CAD device, and the images were randomly selected from the population of patients, then this set of patients may be used to test the performance of the CAD device. The performance, e.g. sensitivity, of the CAD device on these images should be an unbiased estimate of how well the device will perform on the general population of patients[17].

D. Regulation of CAD devices

1. Classification of medical devices

A basic premise of the law is that the degree of device regulation should correlate with the degree of risk posed by the device.[18] Devices are thereby placed into one of three classifications: Class I (low risk)—e.g., stethoscopes; Class II (moderate risk)—e.g., most imaging devices such as CT scanners, 3D reconstruction software; and Class III (high risk)—e.g., pacemakers.

Class I devices are subject to "General Controls." These are baseline requirements that all approved devices must satisfy. However, they are the only requirements that Class I devices must satisfy. General controls include: registering with the FDA and listing all of the devices that a given manufacturer produces; compliance with the adulteration and labeling provisions of the Food Drug and Cosmetic Act (FDCA) and all records and reporting requirements; and the manufacture of devices in accordance with the Quality Systems regulations (i.e., Good Manufacturing Practices).

Class II devices are those for which general controls alone are insufficient to assure safety and effectiveness, and for which there is sufficient information to establish "special controls" to provide such assurance. Special controls can include[19] the promulgation of performance standards, post-market surveillance, patient registries, development and dissemination of guidance documents (including guidance on the submission of clinical data in premarket notification submissions in accordance with section 510(k) of the act), recommendations, and other appropriate actions as the Commissioner deems necessary to provide such assurance.

Class III is the most stringent regulatory category for devices. Class III devices are those for which insufficient information exists to assure safety and effectiveness solely through general or special controls. Class III devices are usually those that support or sustain human life, are of substantial importance in preventing impairment of human health, or which present a potential, unreasonable risk of illness or injury. Other devices may be Class III because: they are used in conjunction with Class III devices; use of the device has a large potential effect on the public health (e.g., due to the particular disease being diagnosed or treated); the scientific concepts

that form the basis for the device are not well-known or established; or because of other factors that influence the determination of safety and effectiveness.

For purposes of device classification,[20] the safety and effectiveness of a device are determined with respect to

- the persons for whose use the device is represented or intended
- the conditions of use prescribed, recommended, or suggested in the labeling of the device
- the probable benefit to health from the use of the device weighed against any probable risk of injury or illness from such use.

2. Regulatory submissions for medical devices

510(k)

Most Class II devices are cleared for marketing via a "510(k)" (named after the section of the statutory law) in which the sponsor demonstrates *substantial equivalence* to a previously approved, legally marketed device (referred to as a *predicate device*). The predicate device is typically a device that has already undergone the 510(k) clearance process but can also be a pre-amendments device (i.e. legally marketed before 1976).[21] A device is substantially equivalent if, in comparison to a predicate it [22]:

- has the *same intended use* as the predicate; and
- has the *same technological characteristics* as the predicate;

or

- has the *same intended use* as the predicate device; and
- has different technological characteristics (such as a significant change in the materials, design, energy source, or other features of the device from those of the predicate device) and the information submitted to FDA:
  - does not raise different (i.e., new) types of questions of safety and effectiveness; and;
  - demonstrates that the device is *as safe and effective* as the legally marketed device.

To help determine the intended use of a medical device that is the subject of a 510(k) submission and to facilitate the comparison of intended use between a proposed device and the predicate, sponsors provide an "Indications for Use" (IFU) statement as part of the 510(k) submission.[23] The IFU statement should include specific indications, clinical settings, define the target population, anatomical sites, etc. The IFU must be consistent with labeling, advertising and instructions for use.[24]

With regard to software devices (particularly devices that are either composed solely of software, such as CAD devices, or devices whose primary effect is essentially entirely software dependent), the term "technological characteristics" refers to the software itself and the specific algorithms (even the specific computer code) that is used. Unless the software code or specific software algorithms in two medical devices are essentially identical, there will be "different technological characteristics." Therefore, 510(k) submissions for devices composed solely of software (or whose effect is strongly or primarily dependent on software) will typically require

information that demonstrates that the device is *as safe and effective* as the legally marketed device when used in accordance with the intended use.

The amount and type of information necessary depends on the complexity of the software and the level of risk associated with the software. For example, basic reporting software may only require straight-forward verification of each function. More complex software, such as CAD devices, may requirement more sophisticated verification and validation data including a clinical trial.

Premarket Approval Application (PMA)

Most Class III devices are approved for marketing via a PMA in which the sponsor provides valid scientific evidence" to provide reasonable assurance that the device is *safe and effective for its intended uses*. Unlike a 510(k) premarket notification, the PMA is not a comparison to other legally marketed devices but must stand on its own to demonstrate the safety and effectiveness of the device for its intended use.

3. Definitions of safety and effectiveness

The FDA defines **safety** as follows[25]

> "There is reasonable assurance that a device is safe when it can be determined, based upon valid scientific evidence, that the probable benefits to health from *use of the device for its intended uses and conditions of use*, when accompanied by adequate directions and warnings against unsafe use, outweigh any probable risks. The valid scientific evidence used to determine the safety of a device must adequately demonstrate the absence of unreasonable risk of illness or injury associated with the use of the device for its intended uses and conditions of use. Among the types of evidence that may be required, when appropriate, to determine that there is reasonable assurance that a device is safe are investigations using laboratory animals, investigations involving human subjects, and nonclinical investigations including in vitro studies."

The FDA defines **effectiveness** as follows[26]

> "There is reasonable assurance that a device is effective when it can be determined, based upon valid scientific evidence, that in a significant portion of the target population, the *use of the device for its intended uses and conditions of use*, when accompanied by adequate directions for use and warnings against unsafe use, will provide *clinically significant results*. The valid scientific evidence used to determine the effectiveness of a device *shall consist principally of well-controlled investigations* as defined in paragraph (f) of this section, unless the Commissioner authorizes reliance upon other valid scientific evidence which the Commissioner has determined is sufficient evidence from which to determine the effectiveness of a device, even in the absence of well-controlled investigations. The Commissioner may make such a determination where the requirement of well-controlled investigations in paragraph (f) of this section is not reasonably applicable to the device."

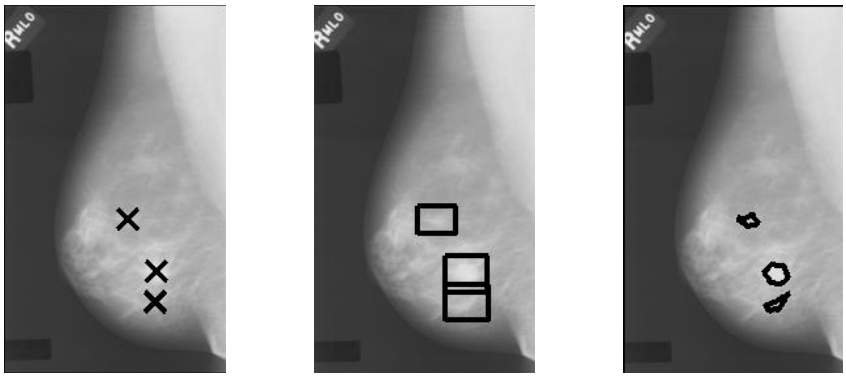
II. General Methodologies for Computer Aided Detection

A. Clinical use of computer aided detection

1. Use of CAD devices in clinical practice

Computer-assisted detection (CAD) devices are used by physicians who interpret medical images to help them identify regions of (or specific findings on) an image that may require them to take a clinical action. The physician makes the decision as to whether to take the clinical action (e.g., obtaining additional imaging, biopsy, etc.), and in many cases, no clinical action will be warranted. Physician performance may be influenced by a variety of factors including the following:

- physical characteristics of the CAD mark:[27] Figure 3 shows an example of a mammography CAD device output with different types of prompts. Physicians may respond differently to different types of marks.



(a) (b) (c)

Figure 3: Examples of different mammography CAD prompts.[28]

- number of CAD marks and the radiologist's knowledge of the sensitivity and specificity of the CAD standalone performance may affect the radiologist's confidence in CAD marks and/or the attention he/she gives to other regions of the mammogram.

CAD may also influence the time required for interpretation in any of the following ways:

- Increase interpretation time if the radiologist conducts both a complete review of the case and then devotes additional time for evaluation of CAD marks (sequential reading mode)

- Keep total interpretation unchanged (i.e., similar to interpretation without CAD) by devoting less time to interpretation without CAD, and the rest of the time to evaluating the CAD marks (sequential or concurrent reading mode)

- Decrease total interpretation time by reducing with and without CAD interpretation (first or concurrent reading mode).

Factors that might influence clinical interpretation should be considered when designing and testing CAD devices.

2. Reading paradigms

Reading paradigms (or modes) specify how a CAD device should be used by the physician when interpreting cases clinically. There are three basic paradigms for CAD implementation:

- First reader mode: The physician reviews only regions or findings marked by the CAD device. Unmarked regions may not be evaluated by the physician. A device of this type has been approved for identifying cervical cytology slides that need no further review (AutoPap, P950009). No CAD device for radiological images with a first read indication for use has been approved or cleared by FDA

- Sequential (or Second) reading mode: The physician first conducts a complete interpretation without the CAD device (unaided read) and then re-conducts an interpretation with the CAD device (aided read). The term *second reader* is often used to describe this reading paradigm. However, the term "second detector" would be more appropriate.

  Various CAD devices (i.e., in mammography, lung) with such indication for use have been approved by the FDA. Some devices intended to be used in a sequential reading mode have been labeled with the "always-never rule," i.e. the physician should always read the radiological image before turning the CAD on, and the physician should never ignore findings detected before turning the CAD on in response to the CAD device not marking the finding.

- Concurrent reading mode: The CAD marks are available at any time and the physician performs a complete interpretation in the presence of CAD marks.

Devices can be tested and labeled for one or more specific reading paradigms.

B. Nonclinical testing of CAD devices

CAD devices are computer software that performs extensive processing and analysis on medical images. Different CAD devices contain different software and utilize different processing, algorithms and features of the image to identify abnormalities. The process of training algorithms or selecting features varies across CAD devices and this influences the performance of these devices. Most processing methods, algorithms, training and selection techniques are known and well described in published literature. There are, however, almost unlimited ways in which these methods can be combined and optimized to make a CAD device.

To fully understand how a particular CAD device works the following would have to be known about that device:

- General Information
  - For what and whom is the device targeted?
  - What imaging modalities are available for use with CAD device?
  - On what kind of system will the CAD be installed?
- Intended Device Usage
  - How will the device be used? How the device is used may affect how the device functions. For example:
    - Will operations and settings be manual (by the physician), semi-automatic, or completely automatic (nothing is under physician control)?
    - What kind of output is generated by the CAD device? Is there user feedback?
- Processing
  - What processing is performed on the medical images? Examples include organ segmentation, filtering, noise reduction, normalization, etc.
- Features
  - Which features were computed or evaluated during algorithm development? Feature information includes
    - the mathematical formulation of the feature
    - the feature description in terms of biology, morphology or geometry
    - the procedure for selecting features
- Classifiers and Models
  - Which models, algorithms, and/or classifiers are utilized in the CAD device (e.g., simple threshold, decision tree, linear classifier, neural network) and how are they optimized? The kind of algorithm and how it was trained will substantially affect how well the CAD device performs and how stable the device is.
- Training Databases
  - Typically data used to train the CAD devices (the training set) will consist of a set of patient cases who have the disease or abnormality of interest and a set of cases who are normal, demonstrating no disease. The training set data will have a strong influence on the behavior of the resultant CAD device. It is important that a training set consist of cases with abnormalities typical of those to be identified in the intended use population of patients. Thus to fully understand a CAD classifier it is necessary to know at least the following about the training database:
    - patient demographic data, e.g., age, ethnicity
    - collection sites and processing methods
    - the number of disease and normal cases
    - the cases by:
      - types of disease,

        - organ characteristics (e.g., breast composition)
        - imaging acquisition hardware and protocol
- Algorithm Stability
    - CAD algorithms can be stable or unstable. If stable, they give substantially similar results when minor changes are made to the algorithm, features, training, or the training databases. If unstable, they will change their output drastically when such minor changes are made. CAD devices, such as those devices approved by the FDA, are regularly modified, making stability an important consideration for how they are regulated. Note, the stability of an algorithm increases as
        - the number of training cases increases
        - the number or dimensionality of initial features decreases
        - the complexity of the CAD decreases
    - The stability of algorithms that are trained using automated or semi-automated methods can be measured by repeatedly retraining the algorithm with substantially different training case sets, and then testing the algorithm on an excluded case test set.

Topic for discussion

G1. To what extent should Sponsors provide algorithm descriptions, training dataset descriptions, standalone performance of the device on the training database, and/or stability analysis of the algorithm to training as part of original CAD submissions or as part of subsequent algorithm updates?

C. Clinical testing of CAD devices

1. Hierarchical model of efficacy

The discussion of CAD testing is a subset of a larger set of issues in medical imaging usually framed in terms of the six-tiered or hierarchical model of efficacy described by Fryback and Thornbury,[29] This hierarchy was further formalized by a report committee.[30] We have reproduced the discussion of this hierarchy from Wagner et al.[35] A streamlined version of this model is given in **Table 1**.

Table 1: The Six-Tiered or Hierarchical Model of Efficacy[29,30,]

Level 1	**Technical efficacy**	Physical performance measurements, preclinical standalone and bench tests
Level 2	**Diagnostic accuracy**	Sensitivity, specificity, ROC curves, and their summary measures
Level 3	**Diagnostic thinking**	Effect of imaging test on clinicians' subjective estimates of diagnostic probabilities, pretest to posttest
Level 4	**Therapeutic efficacy**	Effect of diagnostic imaging or test on therapeutic management of patients
Level 5	**Patient outcome**	Expected value of test information in terms of gains in quality-adjusted life years (QALYs); also, cost per QALY gaine.
Level 6	**Societal efficacy**	Cost-effectiveness and/or cost-benefit analysis from the societal viewpoint

The first tier, technical efficacy, includes measurements of the physical performance of imaging systems, as well as other bench tests and standalone measurements on CADs. The second tier, diagnostic accuracy, includes measurement of sensitivity, specificity, the receiver operating characteristic (ROC) curve and its summary measures (discussed in the following sections). The first two tiers are more commonly the ones on which industry sponsors of imaging technologies focus when seeking entry to the marketplace through the FDA, although sponsors and the FDA are not constrained to these levels. However, depending on the intended use of the device, testing at other levels may be appropriate. The higher level claims regarding efficacy of patient outcome or societal efficacy are more typically the domain of other agencies. Controlled studies at those higher levels are typically much broader in scope—and thus more difficult and expensive— than those at the lower levels.

A general principle of the hierarchical framework is that, for an imaging procedure to be efficacious at a higher level in the hierarchy, it must be efficacious at the lower levels. This is a necessary but not sufficient condition.[30]

2. Types of test and potential biases

CAD devices discussed here for radiological images are intended for use to aid or assist the physician. They are not for use without physician oversight. There are two primary categories of tests for CAD devices:

- *Standalone performance testing*: performance of the device by itself, i.e., Does the device mark the correct locations in the absence of physician interaction?

- *Reader performance testing*: performance of physicians using the device, i.e. Does the physician recognize disease locations when using the CAD device?

These two types of tests provide different types of information. Standalone test results provide a measure of intrinsic functionality of the device. Reader studies measure the impact on physician performance. If, for example, the standalone sensitivity of a device was good, but the device only marked locations that readers found on their own, then the device would be unlikely to improve physician performance. On the other hand, physician performance may be improved by a device with poor standalone sensitivity, if the device marks locations that readers tend to miss, and the readers recognize these locations as important when they review the CAD marks.

Standalone or reader performance testing is subject to many sources of statistical bias that can often be minimized through good study design. Statistical bias is a tendency for a performance estimate in a study to be misaligned with the true performance in the intended use population. Definitions of bias for diagnostic tests are provided below from two well known statistical references:

- The Statistical Evaluation of Medical Tests for Classification and Prediction, Pepe, M.S., Oxford University Press 2003: [31]
    - *Verification Bias*: Not all patients receive gold standard diagnosis;
    - *Spectrum Bias*: The sample studied does not represent the population;
    - *Test Interpretation Bias*: Information is available that can distort the test results; and
    - *Extrapolation Bias*: The conditions or characteristics of populations in the study are different from those in which the test will be applied. Test integrity issues can stem from knowledge of the reference standard influencing test results or vice versa. In a comparison of several tests, knowledge of one test influencing another is also an integrity issue. For example, if a reader knows that an image is from a person diagnosed with the reference standard as having a disease, that reader will likely be more attentive to the image.
- Statistical Methods in Diagnostic Medicine, Zhou, et al., Wiley and Sons, Inc, 2002:[32]
    - *Selection Bias*: The composition of the sample is influenced by external factors, so the study sample is not representative of the target population (pp. (pp. 68-71);
    - *Imperfect Reference (Gold) Standard Bias*: The reference procedure is not 100% perfect (pp. 358-361);
    - *Test-review Bias*: The diagnostic test is evaluated without proper blinding of the results from the gold standard or the competing test ();
    - *Incorporation Bias*: The results from the diagnostic test influence the subsequent clinical workup needed to establish the patient's diagnosis (pp. 73-74);
    - *Diagnostic-Review Bias*: The gold standard is evaluated without proper blinding of the results from the test under study (p.86);

- *Reading-Order Bias*: When comparing two or more tests, the reader's interpretation is affected by his or her memory of the results from the competing test (p.86); and
- *Context Bias*: When the sample prevalence differs greatly from the population prevalence, the reader's interpretations may be affected, resulting in biased estimates of test accuracy (p.88).

Biases in film reading performance studies are reviewed in Brealey[33].

3. Test dataset

A test database is a set of clinical images with which testing is conducted to determine the safety and effectiveness of a CAD device. It must be different from the set of images used to train/develop or validate the algorithm. The test database represents the target population and the target disease condition, and may include cases that may be difficult to diagnose, cases representing relevant confounders (e.g., imaging findings that mimic disease), and cases with imaging characteristics that may independently affect image interpretation (e.g., breast composition). The test database usually includes a clinically relevant spectrum of patient characteristics, imaging hardware and imaging protocols. For example, differences in images characteristics may result when using different dose for a CT acquisition (e.g., low, normal); differences in patient characteristics (e.g., differences in residual stool content may result from different colonic cleansing protocols).

There are various methodologies to obtain an appropriate database for device testing:

- Collection during **field testing**

  Cases are collected and interpreted in real-time during clinical practice without and with the CAD device.

- **Enrichment** for low prevalence disease

  When disease is present in a very small portion of the target population, a large study would be needed to obtain disease cases. A solution is to **enrich** a non-disease database with disease cases at a higher level than would be found in the target population.

- Enrichment for a **stress test**

  A stress test is an enrichment of non-disease cases with disease cases that are specifically chosen to contain imaging findings that are more challenging for radiologists to detect.[34]

Standalone performance testing can be performed with a database that is collected during a field test or enriched for low prevalence disease. The latter is much more commonly used in CAD assessment. For some CAD devices, it may be appropriate to use enrichment for a stress test comparison of CAD performance with another reading modality.

Reader performance testing may be obtained using any of the three databases described above. The number of cases necessary for clinical performance testing will depend on the pre-established performance endpoints selected and the size of the expected effect. Appropriate statistical sample size estimation can be used for the full population as well as important sub-strata that will be considered in the analysis.

Bias can be introduced during case collection because of spectrum bias, context bias, and selection bias:

- Enriching a sample with difficult disease cases (stress test) can tend to depress performance measures relative to the target population. With this form of **spectrum bias**, results may look artificially poor when compared to studies that do not use difficult populations. What a sponsor gains in time and efficiency using an enriched study could possibly be lost in competitive disadvantage due to depressed performance measures. On the other hand, when comparing performance with and without CAD results, the CAD effect on a reader can be enhanced because the CAD may have more opportunities (i.e., more missed disease cases) on which to improve reader performance.

- Enriching a sample with a large number of disease cases can in general change performance relative to the target population. Readers may notice the enrichment relative to their clinical practice and could adapt to it, creating **context bias**. For example, reader might adapt by becoming more aggressive at calling disease.

- Choosing cases known to be easier to detect under a particular modality is an example of **selection bias**. This can be deterred by prospective random sampling.

Once a test dataset has been collected, an important consideration is whether or not that test data can be reused to evaluate subsequent algorithm revisions. The ideal approach is to develop the CAD algorithm, collect test cases, apply the CAD and report the standalone and/or reader performance in an FDA application. This keeps the testing completely isolated from the training process. However, on subsequent algorithm revisions, the sponsor may want to compare performance using the same test cases or an expanded version of this dataset. It is possible that the CAD developer 'learned' something by simply knowing how the original CAD performed on the test data. This could then be used to produce a revised algorithm using this knowledge. Therefore, after testing a CAD once, the test cases are no longer completely isolated from the training. Some incremental knowledge has been gained and 'contaminated' the test set.

If the test database is large and the feedback minimal (e.g, reporting only pass/fail for a predefined performance level), the contamination and subsequent knowledge gain may be quite small. This suggests that it may be possible to reuse test data under appropriate constraints to streamline CAD assessment. The test set could also grow more efficiently because previously used data would not have to be discarded. Clearly, there would need to be limits for this to be a viable approach to evaluating CAD software revisions.

Topic for discussion

G2. What may be appropriate constraints on the reuse of test data in order to balance data integrity and data collection for CAD assessment?

4. Ground truth

Both standalone and reader performance testing require a ground truth determination. Ground truth determination includes whether or not disease is present within a patient as well as the precise location and local extent of disease. Ground truth determination for a cancerous lesion is usually determined by pathologic examination of tissue samples. Ground truth determination of non-cancerous findings can be made using imaging follow-up over an appropriate time interval to demonstrate no change in a finding (e.g., one-year follow-up for mammographic findings; two-year follow-up for lung nodules). Ground truth determination of the location and local extent of disease can be made by having an expert panel review all available clinical information from the patient, including pathology reports, in order to precisely identify and mark both the location and extent of the disease.

The ground truth determination by a panel of experts is susceptible to reader variability. Independent determination by each panel member provides measurement of the variability associated with the ground truth determination and/or allows for a consensus truth definition.

Ground truth determination for a non-cancerous lesion could be determined based on the imaging finding(s) itself (when tissue sampling is either impractical or unavailable). For example, ground truth determination of lung nodules could be made for both the presence of the abnormality as well as location and local extent by interpretation by an expert panel. Such ground truth relies on physicians' interpretations which are subject to variability, i.e., inter- and intra-reader variability. Wagner et al.[35] and Miller et al.[36] described a methodology to account for such variability in ground truth determination.

Ground truth determination is subject to imperfect reference standard bias, workup bias, incorporation bias and verification bias. For example:

- Optical colonoscopy may be considered the reference standard to determine the presence or absence of polyps. However, some studies have shown that optical colonoscopy is an imperfect standard and that some polyps seen on CT may not be visible at optical colonoscopy. Therefore, one could consider a combination of optical colonoscopy and a panel of radiologists reviewing the CT images as the reference standard. A comparison of CT colonography with CAD versus without CAD under the above conditions will suffer from workup bias and/or incorporation bias.

- Screen film mammograms are used in diagnosis for a pool of patients from which cases to enrich a study sample are chosen. If the comparison is being made between CAD on digitized screen film and CAD with FFDM, the study suffers from incorporation bias because the screen film is used in part to enrich the sample.

If ground truth incorporates the device (or reader) being evaluated, the ground truth will tend to result in an overestimate of the true performance of the device (or reader).[37] For example, if ground truth for breast disease is based only on the pathology of biopsies at reader/device identified locations, there may be diseased patients who are not identified, and the sensitivity of the reader may be overstated.

5. Standalone performance testing

Standalone performance is used to indicate the performance of the device in the absence of a reader, i.e. How often does the CAD device mark regions of known abnormalities (true positives) and how often does the CAD device mark regions that do not contain abnormalities (false positives)? Standalone performance is typically measured per abnormality, often called per lesion. These standalone performance measures are important to the user of the device by potentially influencing their confidence in the CAD marks. Standalone testing can also provide information about the effectiveness of the device to mark important subtypes of disease, patient and imaging characteristics, or help determine which patients might benefit most from the CAD. Therefore, standalone performance typically is stratified by important confounding subtypes.

The standalone performance is also a method for comparing new and/or modified CAD devices. Such comparison is better if the same database of patients is used, or the different databases have nearly identical characteristics, including important confounding subtypes. In any case, a testing database should remain separate from the data used for previous development, training or validation of the CAD algorithm. Such comparison also depends on similarity of truthing and scoring methods. Figure 4 shows how standalone performance can be estimated.

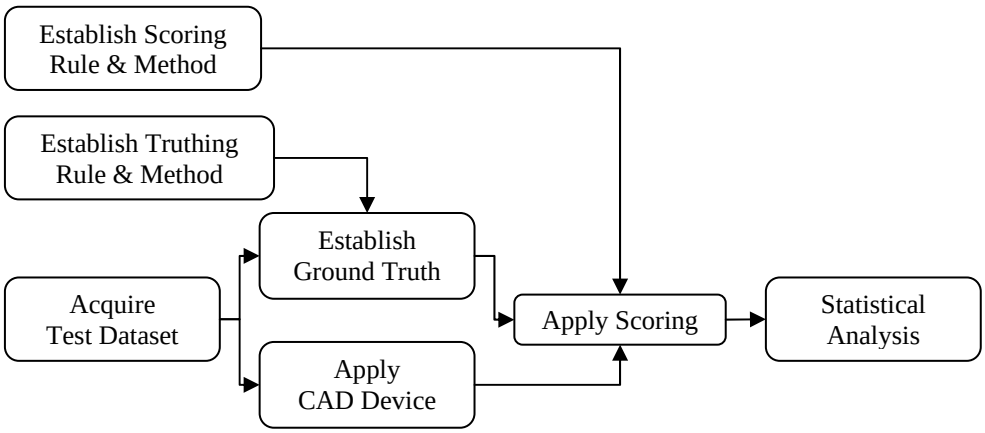


Figure 4: Block diagram of the basic steps for evaluating a CAD device's standalone performance.

Scoring rules and methods

To determine whether a CAD marks the ground truth, a scoring metric must be defined. Different metrics can lead to different standalone performance estimates. Examples of scoring metrics include:

- the ratio of the area/volume overlap between the CAD identified region and the ground truth region to the area of ground truth region
  - $\frac{A_{CAD} \cap A_{truth}}{A_{Truth}}$
- the ratio of the area/volume overlap between the CAD identified region and the ground truth region to the area of the CAD region
  - $\frac{A_{CAD} \cap A_{truth}}{A_{CAD}}$
- distance from the centroid of the CAD identified region to the centroid of the ground truth region
  - $\mathrm{dist}\left(\mathrm{Cent}(R_{CAD}), \mathrm{Cent}(R_{Truth})\right)$
- scoring of CAD marks by a physician (when multiple readers are involved in scoring, an additional procedure or rule for combining their interpretations is needed).

Many other alternative measures are also possible. A clinically relevant threshold is applied to the scoring metric to determine whether or not a CAD mark is a true positive (TP) or a false positive (FP) detection.

Measuring standalone performance

Many measures can be used for CAD device standalone performance, including:

- lesion-based sensitivity and number of FPs per image (or per scan)
- Free Response Receiver Operating Characteristic (FROC) curve (see Section II.D.3.)

CAD device standalone assessment should be measured in accordance with pre-established endpoints. When measuring standalone performance, the types and nature of the abnormalities marked (or not marked) by the CAD device should be consistent with the intended use of the device. Features of standalone performance study design that may impact the interpretation of standalone test results include the following:

- how test database (both diseased and normal cases) represents the intended use population or a challenging subpopulation
- sample size in relation to performance claims
- sample size in relation to important population sub-strata
- method for determining ground truth

- scoring method to identify which CAD marks sufficiently overlap true abnormalities to be considered as TP detections
- study endpoints, e.g. per lesion and per patient sensitivities stratified by lesion type (e.g., breast masses vs. calcifications), number of false positives per abnormal image, and/or number of false positives per normal image

A pilot study may be useful to vet the ground truth and scoring methodologies, to identify study endpoints, and to determine appropriate sample sizes

Repeatability and Reproducibility testing measure

For some CADs, repeatability in the marks displayed is a concern. For example, for breast CADs, a scanner is used to obtain a digital image from a film, and the digitized image is then processed by the CAD. Because the digital image will vary slightly from scan to scan, the CAD output could also vary. Repeated scans can be used to study this variability in a repeatability study. With reproducibility studies, factors such as different scanners, different scanner manufacturers, and different scan operators can be varied to investigate their contributions to total variability in the CAD output.

An important type of reproducibility study evaluates the difference between the CAD outputs from an approved/cleared version and the next generation of the algorithm. A small variation in the marks between the current and next generation device may support approval/clearance of the new CAD without a reader study (see below). The potential benefit of doing a reader study may depend on the kinds of changes that were made to the CAD algorithm and the magnitude of the variability in the CAD outputs.

6. Reader performance testing

CAD devices are intended to mark, highlight, or otherwise bring regions of (or findings on) a patient image to the attention of a physician who determines a clinical action. Reader studies are designed to determine the impact of a CAD device on a physician performance.

Features of reader performance study design that may impact the interpretation of reader performance test results include the following:

- how study population (both diseased and normal cases) represents the intended use population
- sample size in relation to performance claims
- sample size in relation to important disease subtypes
- method for determining ground truth and for accounting for uncertainty in ground truth
- definition of localization of disease by CAD
- how readers represent intended users of the device.

If an approved/cleared CAD device is being modified, additional testing might be necessary when modifications (related to the intended use and/or the device technology) could compromise safety or effectiveness. Recommendations for testing of a modified device could depend on the following:

- the results of the standalone performance assessment of the modified device
- changes to the indications for use or intended use
- changes to the algorithm design
- changes to the functionalities of the PACS or display system on which the algorithm is installed
- use of new technology
- post-market study results for the approved/cleared device
- changes of the labeling

Figure 3 illustrates how reader performance without the device could be compared to reader performance with the device.

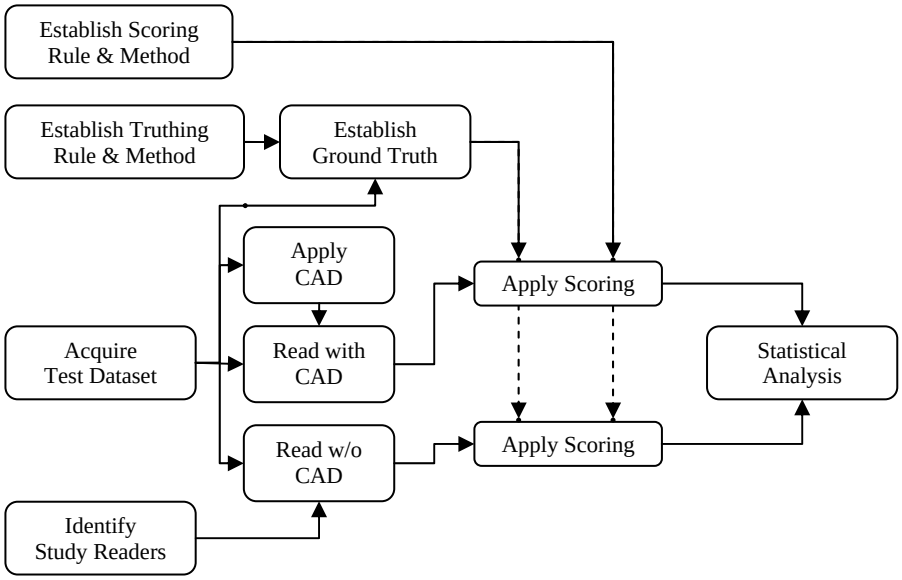


Figure 5: Block diagram of the basic steps for evaluating the reader performance testing of a CAD device.

Study design

Reader performance testing can be done using a field test or ***prospective*** reader study (e.g., randomized controlled trial), or an enriched ***retrospective*** reader study. A field test evaluates a device under real-time and real clinical conditions. A field test may not be practical, e.g., for very low disease prevalence, a field test would have to be huge. Under these circumstances, a retrospective study enriched with disease cases is a

possible surrogate for evaluating a CAD device. A stress test can also be considered, that is, a retrospective study that is enriched with more challenging findings for radiologists to detect that seen in practice. Prospective studies can also be enriched using stratified random sampling[38] (Obuchowski NA and Zhou XH, Prospective studies of diagnostic test accuracy when disease prevalence is low, *Biostatistics* (2002), 3 (4), 477–492.

- *Prospective Study*

In a prospective study design, patients are randomized to two or more reading modalities (e.g., readings with or without the assistance of CAD). The interpretation mode for prospective studies involves a single reader for each case in the course of routine clinical practice. Such studies may necessitate a prohibitive numbers of readers and cases due to reader and case variability as well as low disease prevalence.

Large scale prospective studies are infrequently a condition for premarket clearance/approval of a device. Unlike, for example, mammography, most imaging tests have numerous clinical indications (i.e., many different signs or symptoms), and the images may reveal many different types of findings. Many of these findings are completely unrelated to the imaging findings for which CAD is indicated. Under such circumstances, the relation between the CAD device and other diagnoses may be explored with a prospective study.

- *Retrospective Study*

Multiple Reader Multiple Case (MRMC) Study Design

In a retrospective reader study design, cases are collected prior to image interpretation and are read offline by one or more readers under specific reading conditions (without or with the CAD). Retrospective reader studies have been previously used by companies to support the approval of CAD devices.

In one types of retrospective reader studies, a set of readers interpret a common set of patient images, in each of two competing reading conditions (e.g., readers unaided versus readers aided by CAD) . Such studies are referred to as multiple reader multiple case (MRMC) design. The MRMC design can be “fully-crossed” whereby all of the readers independently read all of the cases. This design offers the most statistical power for a given number of truth-verified cases.

Concerns about retrospective studies include the following:

- The behavior of readers may change, because readers know that they are participating in a study that does not affect patient management.
- With enriched studies readers may notice the enrichment relative to their clinical practice and could adapt accordingly, a form of context bias[48]. For example, some readers might become more aggressive in calling a lesion malignant.

If readers are instructed to read sequentially first unassisted by CAD, and then aided by CAD as a second reader, then the comparison is between an unaided read and the unaided read combined with the second CAD-aided read. Note the unaided read is actually part of both modalities being compared. Concerns with this study design are

- A reader's unaided read could be affected because he knows assistance from CAD is forthcoming. The reader might either overcall the unaided read (competition bias) or undercall the unaided read (satisfaction bias).
- In general, the study design is subject to test-review bias and reading-order bias, as described in Zhou et al.,[48] and can also fall into the general category of studies with integrity problems as discussed in Pepe.[31]

As an attempt to overcome the potential reading bias in the sequential design, an alternative design is to randomize readers to either make or not make the second CAD-aided read, with the randomization revealed only after the initial unaided read is made. In this randomized unpaired design, the comparison between the unaided and CAD-aided reading modalities is a comparison between images. An alternative design that enables comparison of modalities within images is to have two separate reading sessions of the same images, separated by a washout period to erase reader memory of the images, Readers read images once in each reading session. For each image, reading modality is randomized to reading session. This design allows readers to make readings of each image under both modalities, but not know whether the initial unaided read will be followed by a second CAD-aided read.

Although reading behavior in a retrospective study might not be the same as in a clinical setting, the hope is that when comparing reading modalities, relative performance is approximately preserved, even if absolute performance is not, so that CAD effects extrapolate to the clinical setting. .

Warren-Burhenne et al. Study Design:

Early evaluation of mammography CAD devices included two separate studies:

- a retrospective study of the sensitivity of a CAD to detect abnormalities that were "missed" in clinical practice
- a prospective study of the work-up rate when readers use the CAD in clinical practice.

Such retrospective studies have been described in Warren-Burhenne et al.[39] In the retrospective sensitivity study, missed cancers on prior films are assessed by a panel of credentialed readers for visibility and actionability. The fraction of actionable cases detected by CAD is an estimate of the relative reduction in the false negative (FN) rate that could be achieved by a reader using CAD. The estimate is typically weighted according to the fraction of panelists that consider a cancer actionable. The estimate can be an upper bound on the actual relative reduction in FN rate seen in clinical practice, because a reader that initially missed a visible and actionable cancer may still elect to ignore CAD detection of the cancer or continue to misinterpret the location as normal. Further, the panel assessing the cancer for visibility and actionability is typically only shown the location of the missed cancer. That is, they know the cancer was missed, which could affect their evaluation. An alternate setting would be to have the panel evaluate all of the CAD marks together, both true positive and false positive, and blind them to which mark is the missed cancer.

The work-up study has been commonly a prospective study in which radiologists read images without CAD for a period of time in their daily practice. CAD is then installed at their practice and more images are read, this time using CAD. The work-up rate is a measure of the false positive rate because sample size in these studies is usually small

enough and the prevalence of disease is low enough such that that none of the patients are expected to have disease. The observed difference in the work-up rate is attributed to the use of CAD. However, the difference in work-up rate is confounded with the different time periods during which the unaided and CAD-aided readings were taken. For example, the same reading modality could have different work-up rates in the two time periods simply because of differences in the mix of patients who were seen in those periods. This confounding could result in an illogical estimate of the difference in work-up rate. For example, a decrease in work-up rate could be observed, even for a second reader CAD where a decrease in work-up rate is impossible (i.e., the physician is instructed to never eliminate biopsy commitments from an initial CAD-unaided reading when performing a second CAD-aided reading).

A fundamental limitation of the Warren-Burhenne et al. study design is that the reduction in FN rate and the increase in work-up rate are not being evaluated in the same study. and therefore are not likely to be comparable. Valid comparison of the FN rate (e.g., 1 – sensitivity) and the work up rate (e.g., 1 – specificity) is essential to assess the trade-off in these measures that results from using CAD. By not evaluating the FN rate and the work-up rate in the same study, results from a Warren-Burhenne designed study can be difficult to interpret statistically. Furthermore, as discussed above, the design is also problematic because its goal is to estimate the "potential" effect on the FN rate, not the actual effect.

The Warren-Burhenne study design was used in the past to evaluate CAD devices. FDA is seeking panel input as to which study designs are appropriate.

Reader selection

Retrospective reader studies have typically used between 7-12 readers to read images. Participant readers for the study should be representative of the intended user. Factors that may influence reader performance include experience and expertise. Their selection may introduce bias similar to that in acquiring the test data. Care should be taken to ensure that the reader sample represents the population of readers who will use the device in practice as well as representing the proper experience level.

Training

Reader performance testing depends on the complexity of using and understanding the CAD device. To avoid inclusion of bias due to learning effect and adjustment to use of the device during the study, physicians can be trained prior to initiating a pivotal study. Additionally, some statistical analysis such as the ROC and FROC will require physicians to supplement the diagnostic report with a detection and/or diagnostic confidence level or level of suspicious rating on a scale (e.g., 10-point scale, 100-point scale) which is not used in the current clinical practice. Such rating may not be easy for physicians to use since most diagnostic decisions are binary or at best ordinal relative to patient management (e.g., surgery/biopsy/follow-up/normal). Training may allow physicians to adapt to this unconventional rating. Prior to initiating the study, the physician's understanding of the rating scale should be evaluated. There are circumstances where a rating scale may not be practical to use or may be irrelevant, because such ratings are inconsistent with the actual practice of medicine. In addition, it may be difficult to train physicians to adapt their interpretation to some rating scales. An example of this can be found in Petrick et al.[40] If the rating is unusable, alternates would

be to redefine the rating scale or use a different approach for the statistical analysis. This as well as all elements described above could be determined with a pilot study.

D. Primary concepts of statistical analysis

1. Study endpoints

Several endpoints exist to measure the results obtained from use of a CAD device. Below are some measures of detection/diagnostic accuracy:

- **Sensitivity** (Se) is defined as the probability that a test indicates a condition when the condition is present and **specificity** (Sp) is the probability that a test does not indicate a condition when the condition is not present. These two measures are considered together in order to understand how the test should behave with both affected (diseased) and unaffected (healthy) specimens. A test that perfectly discriminates will have Se=1 and Sp=1. If sensitivity and specificity sum to one, the test is useless.
    - Se= TP/(TP+FN)
    - Sp= TN/(TN+FP)

- Two other diagnostic accuracy measures derived from the table are positive predictive value (PPV) and negative predictive value (NPV). Positive predictive value is the probability that a condition is present given a positive test result, and negative predictive value is the probability that the condition is not present given a negative test result.
    - PPV= TP/(TP+FP)
    - NPV= TN/(TN+FN)

- Single dimension diagnostic metrics measuring overall agreement and Cohen's Kappa can be difficult to interpret and are affected by disease prevalence. For these reasons, the FDA discourages their standalone use. Please see: Guidance for Industry and FDA Staff; Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests, available at: http://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfGGP/Search.cfm.

- The receiver operating characteristic (ROC) curve is a curve displaying Se (or true positive fraction (TPF)) as a function of Sp (or false positive fraction (FPF)) when the threshold setting for the test is varied over the complete range of possible test scores or rates.[41] Display of a full ROC curve addresses the wide range of thresholds over which readers may operate. A summary measure of the performance is derived by using the area under the ROC curve (AUC) or partial area under the ROC curve (PAUC). The partial area requires specification of a clinically relevant range for assessment.

2. ROC analysis

ROC methods consider all sensitivities at all possible specificities. An ROC curve is a plot of all sensitivities at all possible specificities. It is a complete summary of diagnostic performance of a device or a physician.

Per patient ROC analysis requires binary truth at the patient level. That is, all cases must be known to be either normal or abnormal (diseased). ROC analysis does not account for multiple diagnoses or marks per patient. In the case of CAD detection devices that do mark multiple locations, some method or system must be used to determine what single diagnosis should be assigned to the patient.

Status A: Abnormal, N: normal	N		A	N		A	N		A	A	A
Ordered ranking by CAD or physician	**1**		**2**	**3**		**4**	**5**		**6**	**7**	**8**
Example thresholds		**B**			**C**			**D**			
True Positive Fraction (TPF)		**1**			$^{4}/_{5}$			$^{3}/_{5}$			
False Positive Fraction (FPF)		$^{2}/_{3}$			$^{1}/_{3}$			**0**			

In ROC analysis a physician can be requested to rank a set of cases from most normal or least likely to contain disease to most obviously diseased. An example is given in the table below. A radiologist ordered eight cases from 1 to 8 where cases that were more suspicious of disease were given higher rankings. Of these eight cases five cases were known to be diseased or abnormal (A) and three cases known to be normal (N).

A physician could decide to send his three highest ranked cases for further diagnostic tests, or he could send the five highest ranked cases. The physician could decide to use any such threshold that he chooses. Three such example thresholds, B, C, and D, are given in the table. We can measure the fraction of abnormal cases that are on the correct (right-hand) side of each threshold (true positive fraction or Se) and the fraction of normal cases to the right of each threshold (FPF or 1-Sp). The TPF and FPF for thresholds B, C, and D are given in the table.

A plot of all possible (FPF, TPF) pairs at all possible thresholds is called an ROC curve. An empirical ROC curve constructed from the data given in the table is shown in Figure 6. Each normal case is represented by a horizontal line segment. Each abnormal case is represented by a vertical line segment. These line segments appear in the order given by the radiologists' ranking and trace out all possible FPF and TPF values. The FPF and TPF values at example thresholds B, C, and D are labeled in Figure 6.

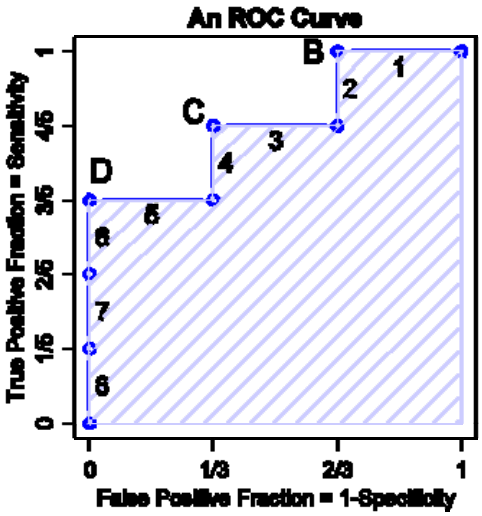


Figure 6: Plot of an ROC and the corresponding area under the ROC curve (AUC) estimate for this curve.

Physicians who are better at diagnosing disease may be able to separate all abnormal cases (A) from all the normal cases (N). In that situation all vertical line segments (true positive cases) will be together on the left of the ROC plot and all horizontal line segments (false positive cases) will be across the top. Such a ROC curve is marked with black circles in Figure 7.

Physicians who are worse at diagnosing disease will have the abnormal and normal cases more mixed together, resulting in a lower ROC curve. In Figure 7 the blue "+" ROC curve is worse than the black circle ROC curve or the red triangle curve. At every possible specificity, the blue "+" curve demonstrates a lower sensitivity than those other curves.

Remember that the physician who ranked the eight cases can choose to send any number of the patients to further diagnostic testing, or short-term follow-up. Which set of patients the physician chooses determines the reader's sensitivity and specificity and this (Se, Sp) or (TPF, FPF) pair falls somewhere on the ROC curve. The point on the ROC curve that the physician chooses is his "operating point". Points B, C, and D in Figure 6 are such possible operating point choices.

Diagnostic or detection aiding devices that force or encourage the physician to change his operating point while remaining on his ROC curve are generally not considered informative, because such a change is a choice that the physician could make on his own. The physician can choose to send more patients for further work up without a device.

For example, a study may indicate that a clinician can make diagnoses with a specificity of 1 and a sensitivity of 3/5, as marked by D in Figure 6. If that clinician uses a diagnostic aid such as a CAD device in a subsequent study, his specificity may decrease to 1/3 and his sensitivity may increase to 1, as marked by B. In general this diagnostic device is not considered informative, because the clinician could have chosen to send the exact same cases (cases ranked 2-5) on to further evaluation without using the CAD device. Similarly if a clinician operated at point X in Figure 7, and with the use of a CAD

he instead operated at point Y, then that CAD device would not be considered informative, because both points are on the same blue "+" ROC curve.

Devices that help the physician better separate abnormal cases (A) from the normal cases (N) will raise the physician's ROC curve and are generally considered informative. For example, if a clinician operated at point X in Figure 7, and with the use of a CAD he instead operated at point Z, then that CAD device could be considered informative, because the CAD enabled the physician to operate on the higher red triangle ROC curve. In this case the CAD enabled the physician to send a different set of additional cases for further work up, a set that contained a larger increase in abnormal cases compared to normal cases than the physician could have by moving up the ROC curve.

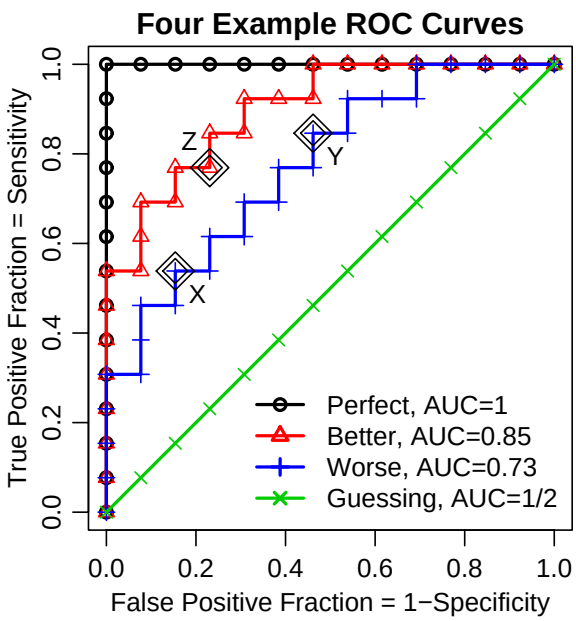


Figure 7: This figure shows four ROC curves. The black circle ROC curve demonstrates perfect diagnostic performance. All vertical line segments (representing diseased cases) are separated from all horizontal line segments which represent normal cases. The red triangle ROC curve shows good, but not perfect diagnostic performance. Here some vertical segments are in mixed order with horizontal ones. In the blue "+" ROC curve demonstrates worse diagnostic performance. The green line indicates an equal likelihood of diseased or normal cases regardless of ranking. This is guessing.

The area under the ROC curve (AUC) is one of the summary measures of diagnostic performance. AUC is TPF averaged over all possible false positive values. The greater the AUC value, the better the diagnostic performance. In our example, the measured AUC is $(3+4+5) \times {}^{1}/_{5} \times {}^{1}/_{3} = {}^{12}/_{15}$.

In an experiment where the reader perfectly separates normal (N) from abnormals, the AUC will be 1 (black curve in Figure 7). In a case where the reader of CAD is simply guessing (i.e., no diagnostic utility), the true positive rate will, on average, equal the false positive rate, and the AUC will be around ½ (green line in Figure 7). The AUC is

equivalent to the Mann-Whitney-Wilcoxon statistic. It is the probability that given a random abnormal case and a random normal case, the physician or device can correctly diagnose which is which.

Here we have presented nonparametric ROC examples. ROC data can also be approximated by curves that assume an underlying distribution of the data (or a transformation thereof). These curves are smooth and can be described by a small number of parameters. They are parametric fits to the data. Popular simple parametric models include the two normal model and the two exponential model. Other models include the contaminated binormal model, the bigamma model, and the PROPROC model. For more information on parametric models see the references by Wagner or Metz.

An alternative summary measure of diagnostic performance is the **partial area under the ROC curve (PAUC)**. The PAUC is TPF averaged over a range of false positive values (e.g., FPF from 0 to 0.3). It requires using at least one clinically relevant FPF threshold. Many continue to use the AUC because of the complexity (even the arbitrariness) of selecting a particular threshold using any amount of limited experience with the utilities especially in modest studies. Because AUC and PAUC average over a range of TPFs, they offer the possibility of greater statistical power (i.e., tighter error bars) in an analysis than use of a single TPF. This is another reason for the popularity of these measures, beyond the desire to avoid a specific but arbitrary threshold.

It is challenging to link the AUC measure to clinical relevance. Furthermore, ROC does not take into account the location uncertainty in patient-based analysis. An alternative is to segment the organ of interest into clinically relevant smaller units (i.e., regions) or, smaller anatomical units, which are clearly defined and visible on the images. Another alternative could be to set a primary endpoint with two hypotheses (co-primary endpoints) to be tested, (i.e., one on the ROC and one on the location percent correct between reader/device result and the ground truth).

3. FROC analysis

Free Response ROC (FROC) is similar to ROC, however it allows for the possibility that each patient may be diagnosed at multiple anatomical locations. FROC does not assume that each patient is diagnosed entirely as normal or abnormal. For example, CAD devices regularly create multiple marks per patient because there may be multiple locations in the image that are indicative of disease. The majority of these marks may be false positive marks. Though there are multiple diagnoses per patient, a patient is still considered a single sampling unit.

Like ROC, FROC compares all true positive rates against all false positive rates. In the case of FROC, patients may have multiple false positive diagnoses. Figure 8 shows an example of an FROC curve.

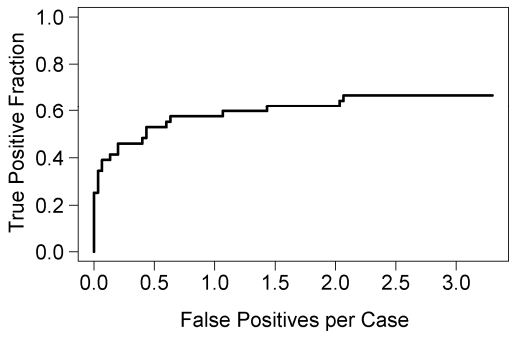


Figure 8: Example of a free-response ROC (FROC) curve. FROC is typically used to plot the performance of CAD devices because it accounts for correct localization of abnormalities within a patient.

There is almost no limit on the number of false marks that could be generated on an image. In general, CAD devices limit the number of marks per patient that are generated. This limit differs from device to device. Frequently, FROC analysis or comparing different devices using FROC requires some sort of artificial truncation of the data, e.g., to enable comparison on the same range of the number of false positives per case.

As with ROC different metrics can be constructed using FROC data. Different performance metrics and analysis methods have been developed to handle such data. They include the following:

- Cutpoint or operating point analysis, i.e., sensitivity, specificity or number of false positives per case.

- Alternative FROC (AFROC) plots the fraction of actual target locations reported against the fraction of images with at least one false-positive at the same rating or threshold criterion (a so-called false-positive image).[42, 43]

    - This plot has unit area so that the area under the AFROC curve is a viable endpoint. However, simplistic AFROC analysis makes independence assumptions that may not be met in practice.[44]

- Jackknife AFROC (JAFROC) is a multiple-reader multiple-case (MRMC) solution to the AFROC problem using a jackknife re-sampling approach that respects the correlation structure in the images. For scorekeeping, they include all lesion ratings on abnormal images (unmarked lesions receive the lowest or default rating, and false-positives on lesion-containing images do not contribute); the summary figure of merit is defined as the probability that a lesion rating on an abnormal image exceeds all false-positive ratings on a normal image. Under these conditions, this paradigm successfully passes a rigorous statistical validation test.[35,45]

The definition of TP and FP CAD marks should be consistent with the intended use of the device. For example, if the device is intended to detect all potential abnormalities (e.g., benign and malignant), then a consistent definition of a true positive CAD mark

should be the "marking" of any abnormality. In this situation, it may be important to stratify per lesion type, because a device could be found to help for one disease type (e.g., benign lesion) that is not as clinically as important as some other types (cancer lesion). When the device is intended to detect only a subset of potential abnormalities (e.g., only those lesions with certain imaging features), then a true or false CAD mark should be defined accordingly.

E. Additional statistical issues

An FDA statistical guidance on diagnostic devices is "Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests", 2007, (http://www.fda.gov/cdrh/osb/guidance/1620.pdf). Although not specific to CAD devices, many of the principles outlined there apply to CAD devices.

1. Comparator for reader studies

In reader studies, the control (comparison) group for CAD aided reading has typically been unaided reading by a single reader. This control is essentially a no-treatment arm. Alternative controls include the following:

- *Unaided double reading by the same reader*: The reader is asked to read once, and subsequently, to look again, to mimic CAD assistance. The reader could be instructed, for example, to look again at specific areas that may have been overlooked the first time. This control can be used to investigate whether the improvement in performance when using CAD is due to taking a second look, regardless of the markings made by CAD.

- *Unaided double reading by two readers*: Unaided reads are made independently by two readers. The findings are combined. This control might be appropriate if the CAD is supposed to be as good as a second reader.

- *Reading aided with a sham CAD*: If feasible, a sham CAD randomly places the same number of marks on each image as the CAD did. (A possible, but, as far as we are aware of, uninvestigated way to create a sham CAD is by lowering the CAD threshold so that many more false positives are called, and then randomly selecting from those.) Comparison with a sham CAD control would eliminate a potential placebo effect when using the CAD. The placebo effect cancels because the reader is blinded to whether he is using the real CAD or the sham CAD. Preliminary investigation of a sham-like CAD was found to decrease sensitivity compared with unaided reading.[46]

Topic for discussion
G3.What are appropriate controls for reader studies?

2. Sample size for retrospective reader studies

In general sample size is determined using assumptions of the data variability, a given desired effect size and acceptable probabilities of rejecting the null hypothesis when it is true (alpha) and not rejecting the null hypothesis when it is false (beta). Sample size calculation is complicated for the mixed models used in MRMC studies because of the many variances or correlations that have to be assumed. Hillis and Berbaum[47] describe how to do power and sample size calculations for the popular Dorfman-Berbaum-Metz (DBM) model.[49] Sample size calculations for the method of Obuchowski can be found in Zhou, et al.[48] The calculations will provide a sample size to provide adequate power under the given assumptions. Reviewing three Summaries of Safety and Effectiveness Data for mammography devices (e.g., P990066, P010017, and P010025), MRMC sample size is on the order of 200 to 600 cases with approximately 50 to 150 cancers among those cases. Five to twelve readers participated in those studies. Since statistically significant results can be obtained using such a small number of readers care has to be taken in sampling the reader population so that it is properly represented.

3. Statistical analysis of retrospective reader studies

Analysis of fully-crossed multi-reader, multi-case study (MRMC) designs

In a common type of retrospective reader study design, multiple readers and cases (e.g., images) are "fully crossed", that is, every reader reads every case. This design can be very efficient for detecting reading modality effects on performance. For this design, the statistical analysis needs to respect correlation among repeated readings of the same case, correlation among readings of cases by the same reader, and correlation among repeated observations per patient (e.g., two breasts per patient). The analysis also needs to recognize that cases and readers are random samples of case and reader populations, respectively, i.e., that they have random effects. Investigations have been made of the suitability of particular methods of analysis[126]. Common statistical methods for analysis include random effects modeling of jackknife pseudovalues of area under the ROC curve[49] and clustered ROC analysis[50]. Another possible method of analysis is to use bootstrapping[51]. Other possibilities are generalized estimating equations[52] and analysis of variance (or deviance) according to study design.[53]

Statistical analysis by weighting observations

CAD reader performance studies may include multiple important clinical factors. For example, in studies of CADs for mammography, the following clinical factors may be important: BI-RADS scores, breast density, lesion size, age, type of malignancy (mass vs. calcifications), and type of non-malignancy (benign lesion present vs. not). The distribution of clinical variables is likely to vary from study to study, limiting the comparability across studies of the observed CAD performance. A possible but as yet unutilized approach for mitigating this lack of comparability is to standardize the statistical analysis by weighting observations according to a designated standard distribution of the clinical variables. For example, a variant of direct standardization[54] was used to standardize an analysis of a permanent contraceptive device to the age distribution in a large study of such procedures[55].

Topic for discussion

G4. Please comment on the appropriateness of using standardized weighted analysis as a primary or secondary analysis of a CAD study.

4. Units of analysis

When assessing the of a reader/CAD performance, several levels or units of analysis can be considered:

- per patient
- per region (image divided into regions)
- per lesion (locations identified by reader for biopsy)

Per patient analysis is often a primary analysis because detecting more diseased patients can be at least as important as detecting more disease within a patient. This can be especially true if treatment of disease is not localized. Per region and per lesion analysis can also be important, especially if treatment is localized.

For each possible experimental unit, statistical analysis can be made with regard to the reader binary result: true positive (TP), false positive (FP), true negative (TN), or false negative (FN). For example, a radiologist using a mammography CAD may score each identified location for probability of malignant disease, but these can be dichotomized via a threshold to obtain one of the four binary results. As another example, because of a lack of training in scoring methodology a physician using a colon CAD may only provide binary results regarding identified locations of possibly actionable polyps. In this section, analysis of reader binary results is emphasized.

Per patient analysis

When the unit of analysis is the patient, the definition of reader result is not unique. The per patient reader result is a summary of the reader identified locations for biopsy and the locations of any abnormalities. This summary can be defined in many ways.

Commonly, per patient endpoints (per patient sensitivity, per patient specificity, per patient ROC analysis are defined such that correct identification of the location of the abnormality is not required for a TP to obtain (Section C.1). That is, in a patient confirmed to have an abnormality, a reader that identified locations for biopsy is often credited with a true positive result regardless of whether any of those locations coincide with the location of the abnormality.

When evaluating a reader's binary decisions, the following cases can be identified:

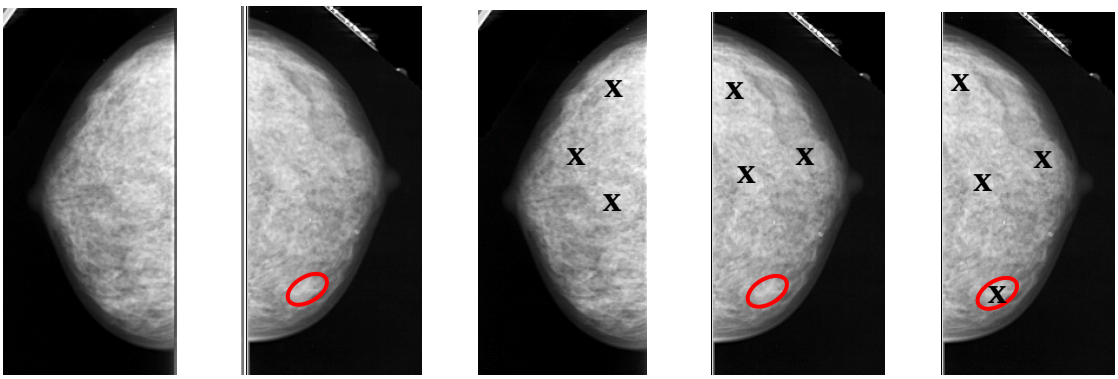


Figure 9: Figure showing possible results for per-patient or per-region analysis for a single region. The red oval represents the true location of an abnormality. The black x's represent reader/CAD identified locations. See text for full description of each plot.

A. Reader does not identify any lesions as abnormal in a patient confirmed to not have any abnormalities.

B. Reader does not identify any lesions as abnormal in a patient confirmed to have an abnormality

C. Reader identifies one or more lesions as abnormal in a patient confirmed to not have any abnormalities.

D. Reader identifies one or more lesions as abnormal in a patient confirmed to have an abnormality, but none of the locations coincide with the abnormality.

E. Reader identifies one or more lesions as abnormal in a patient confirmed to have an abnormality, and some of the locations coincide with the abnormality.

When readers score locations for probability of disease, two common per patient analyses are ROC analysis and FROC analysis. In ROC analysis, the location of disease is typically ignored. That is, the reader is credited with true positive for a disease patient even when none of the locations identified by the reader coincide with the disease location. This approach affects the estimation of per patient sensitivity (proportion of true positives among disease patients). In FROC analysis, disease location is considered in that per lesion sensitivity (not per patient sensitivity) and plotted against the average number of per lesion false positives per image over all possible thresholds in the score.

When scores are not available, ROC and FROC analysis are not possible. If a pure per patient reader binary result is desirable, then the result could be defined in several ways. We bring attention to six possible paradigms that could be considered for defining a binary per patient reader result:

a. *Ignore location of abnormality*: The per patient diagnostic test results are then TN, FN, FP, TP, and TP for A, B, C, D, and E, respectively.

b. *Consider location, interpret D as FP*: D is interpreted as FP because none of the lesions identified are abnormal. The per patient diagnostic test results are then TN, FN, FP, FP, and TP for A, B, C, D, and E, respectively.

c. *Consider location, interpret D as FN*: D is interpreted as FN because the abnormality was not caught by any of the identified lesions. The per patient diagnostic test results are then TN, FN, FP, FN, and TP for A, B, C, D, and E, respectively.

d. *Consider location, interpret D as both FP and FN* (by reasoning given above): The per patient diagnostic test results are then TN, FN, FP, FP/FN, and TP for A, B, C, D, and E, respectively, with D used twice in the analysis to cover both results.

e. *Consider location, interpret D as an undefined result (UR)*: The per patient diagnostic test results are then TN, FN, FP, UR, and TP for A, B, C, D, and E, respectively. Patients with UR results are excluded.

f. *Define test result as positive only if an abnormality is found*: From the point of view of the patient, she is positive by the diagnostic test only if she is managed as if she had an abnormality. The per patient diagnostic test results are then TN, FN, TN, FN, TP for A, B, C, D, and E, respectively. Note an FP is not possible. Consequently, specificity = 1.0, PPV = 1.0, and ROC analysis is not possible.

Table 2: Paradigms for defining per patient reader results for cases A-E

	TEST	A	B	C	D	E
A	Ignore Location	**TN**	**FN**	**FP**	**TP**	**TP**
B	Locate, Interpret D as FP	**TN**	**FN**	**FP**	**FP**	**TP**
C	Locate, Interpret D as FN	**TN**	**FN**	**FP**	**FN**	**TP**
D	Locate, Interpret D as FP&FN	**TN**	**FN**	**FP**	**FP & FN**	**TP**
E	Locate, Interpret D as UR	**TN**	**FN**	**FP**	**UR**	**TP**
F	Test + only if Abnormality Found	**TN**	**FN**	**TN**	**FN**	**TP**

Virtually all clinical actions based on imaging tests are location dependent. However, taking location into account requires a localization criterion as in paradigms b-f.

When defining per patient reader binary results on the basis of actionability or patient management, paradigm a, ignoring disease location, is used. Although a reader may call for more work-up on a disease patient for the “wrong” reason (e.g., a per lesion false positive), the further work-up could still result in discovery of the disease, mitigating the false negative error of missing the disease initially.

Topic for discussion

G5. For per patient analysis, please comment on whether different summaries of per patient results should be considered to assess the robustness of per patient conclusions.

Per region analysis

As with per patient analysis, per region analysis requires a summary of the results within each region. Commonly, the per region reader result is defined such that correct identification of the location of an abnormality within the region is not required for a TP to obtain (paradigm "*a*" above). This type of per region analysis is sometimes used in lieu of per patient analysis because it accounts for correct location of an abnormality up to the precision (size) of the region. The reasoning is if the region is small enough, then identification of a region with an abnormality can be regarded as correct identification of the location of that abnormality for all intents and purposes in most cases. Still, per region analysis may not be an entirely satisfactory substitute for per patient analysis if quantifying results on a per patient basis (e.g., proportion of abnormal patients detected) is important.

In per region analysis, the standard error and confidence interval on a performance estimate needs to account for correlation among the results on the multiple regions within a patient. Clustered data analysis[50] and patient-based resampling are statistical methods that account for this correlation.

Per lesion analysis

The limiting case of per region analysis is per lesion analysis is where the region is shrunk to the size of a location identified by a reader for biopsy. Performance can be assessed by comparing locations identified by the reader to locations unidentified by the reader for presence or absence of an abnormality. Per location analysis is challenging because lesions vary in size, the number of locations that the reader did not identify for biopsy is difficult to define, and without patient follow-up ground truth may only be available for biopsied locations. Under these circumstances, per lesion positive predictive value (PPV) might be computable, but per lesion negative predictive value (NPV), per lesion sensitivity, per lesion specificity, and the per lesion ROC curve are typically not computable.

Table 3: 2x2 tables of per lesion results comparing a reading paradigm using CAD (with CAD) to the unaided reading paradigm (without CAD).

		Abnormality absent CAD assisted read		*Abnormality present* CAD assisted read	
		–	+	–	+
Unassisted	–	**[a]**†	**b**	**[e]**†	**f**
Read	+	**C**	**d**	**G**	**h**

†Count is bracketed to denote that it is undefined

When per lesion reader results are dichotomized as either did identify or did not identify a location as suspicious for abnormality, and these locations are ground-truthed, the data comparing unaided readings with CAD-aided readings can be summarized as two 2 x 2 tables (Table 3). The counts *a* and *b* are bracketed to denote that they are undefined, due to the difficulty in counting the number (*a*+*b*) of unidentified locations and due to lack of ground truth for these locations. As a result, per lesion sensitivity and specificity cannot be computed for the two reading modalities. However, the ratio of sensitivities can be computed as (f+h)/(g+h) and the ratio of false positive rates (1 – specificities) can be computed as (b+d)/(c+d).[56]

One can show that CAD-assisted reading is superior to CAD-unassisted reading if the following are true:

- $Se_A/Se_S > 1$, that is, the per lesion sensitivity ratio of CAD-aided to standard unaided reading is greater than 1
- $Se_A/Se_S > (1-Sp_A)/(1-Sp_S)$, that is, the per lesion sensitivity ratio is greater than the per lesion false positive rate ratio (Sp denotes specificity).

When these two conditions hold, CAD-aided reading is “superior” to standard unaided reading in that both predictive values, positive and negative, are higher for CAD-aided reading than for CAD-unaided reading) [57]. These two conditions are analyzable because the ratios Se_A/Se_S and $(1-Sp_A)/(1-Sp_S)$ observed in Table 3 are not functions of the unknown counts [a] and [e].

III. Mammography CAD devices

A. How mammography CAD devices are used

1. Screening for breast cancer

Breast cancer is the most commonly diagnosed cancer and the second leading cause of cancer-related death among women in the United States. In 2002, at least 182,125 women in the United States had a diagnosis of invasive breast cancer, and 41,514 died from the disease.[58] In 2004, approximately 58% of women in the United States aged 40 years or older reported having had a mammogram within one year.[59] According to FDA statistics as reported by certified mammography facilities in the United States, as of January 1, 2008, there were 8,859 certified mammography facilities, 13,610 accredited mammography machines, performing approximately 36 million annual mammography procedures.[60] Approximately 80% of all mammography examinations are performed for screening.[61]

More than 500,000 women have been entered into randomized trials to investigate the effect of screening mammography on breast cancer mortality. A reduction of the breast cancer mortality on the order of 20–30% has been demonstrated in women who were aged 50 years and older at entry into the trials.[62] Although 25% of all diagnosed cases are among women younger than 50 years of age[63], there remains controversy about the effect of screening women aged 40–50.[64] There is uniform agreement in the recommendations of professional societies that patients who are 50 years or older should undergo yearly screening mammography.

There is variability in the recommendations of professional societies about the need for and/or the time intervals of screening mammography for patients in the 40 – 50 age range: The American Cancer Society[65] and the American College of Radiology[66] recommend yearly screening mammography. The American College of Obstetricians and Gynecologists recommends screening mammography every 1-2 years.[67] The American College of Physicians does not recommend any particular time interval but does recommend that “clinicians should periodically perform individualized assessment of risk for breast cancer to help guide decisions about screening mammography.”[68] The U.S. Preventive Services Task Force (USPSTF) recommends that “women 40 years of age and older consider having screening mammography every 1 to 2 years”[69] The American Academy of Family Physicians recommend screening for breast cancer with mammography every 1-2 years after counseling by their family physician regarding the potential risks and benefits of the procedure.[70] However, an editorial published in June of 2007 in the journal American Family Physician stated that “the balance between benefits and harms is borderline for women in their 40’s [and these women] have to put up with many false-positive tests and some degree of over-diagnosis and over-treatment.”[71]

Recent literature addresses the issue of false positive mammograms (outside of the context of CAD). False-positive (FP) mammograms greatly outnumber actual breast cancers found (approximately 10% of patients who undergo screening mammography will be recalled for diagnostic mammography while approximately 0.2% to 0.8% of patients who undergo screening

mammography have breast cancer). For those patients who are recalled for diagnostic mammography, the FP rates are: 95% for abnormal mammographic interpretations of all types (i.e., other than BI-RADS 1 or 2); 75% when biopsy is recommended; and 66% when biopsy is performed.[72] Nearly 50 percent of all women will have at least one false-positive mammogram over 10 years of screening.[73] False-positive mammograms can cause unnecessary anxiety, increased dose exposure, biopsy, and complications associated with biopsy Overdiagnosis occurs most commonly with ductal carcinoma in situ (DCIS), a lesion that is usually detected by mammography.[74] It is estimated that less than 50% of DCIS lesions progress to invasive breast cancer, yet everyone with DCIS is treated for cancer.[75] This means that overdiagnosis can lead to overtreatment in a considerable number of women. Some studies estimate that about 10 percent of all breast cancers diagnosed by screening represent overdiagnosis.[76]

2. Patient characteristics in a screening populations

The clinical, mammographic and pathologic characteristics of patients who undergo screening mammography in the United States are well-known from large published clinical trials and publicly available databases. The most relevant characteristics include the following:

- **Tumor size**
  Larger tumors are more readily identified and characterized on mammography as compared to smaller tumors. Prior published studies in patients undergoing screening mammography have demonstrated a mean tumor size substantially less than 2.0 cm. For example, in a study by Sickles[77] examining the mammographic features of 300 consecutive non-palpable breast cancers identified on screening mammography, 61% of the cancers were less than or equal to 1.0 cm in size and only 7 of the 300 cancers exceeded 2.0 cm in size. The Breast Cancer Surveillance Consortium (BCSC) database reveals the following distribution of size for invasive cancers: 1-5 mm (10.5%), 6-10 mm (25.8%), 11-15 mm (26.2%), 16-20 mm (14.9%), and > 20 mm (22.7%); 51.9% of tumors were "minimal," where a minimal cancer is defined as DCIS, or, invasive cancer ≤ 10 mm (of the 51.9% of minimal cancers, 15.6% were DCIS).

- **Breast density**
  The ACR BI-RADS system defines four types of breast density:
  - <25% fibroglandular tissue (i.e., almost entirely fatty)
  - 25%–50% fibroglandular tissue (i.e., scattered fibroglandular densities)
  - 51%–75% fibroglandular tissue (i.e., heterogeneously dense)
  - >75% fibroglandular tissue (i.e., extremely dense category).[78]

  The relationship of breast composition to cancer detection is well-known.[79] In general, greater breast density is associated with decreased sensitivity for breast cancer detection and higher incidence of interval development of breast cancer following a negative mammogram. Dense breast tissue can obscure mass lesions and microcalcifications and can make evaluation of imaging features of detected lesions more difficult and less reliable. A published series by Tice et al.[80] that included 81,777 patients undergoing screening mammography revealed the following distribution of breast composition: 10% almost entirely fat; 45% scattered fibroglandular densities; 38% heterogeneously dense; and 7% extremely dense. Similarly, Pisano et al.[81] found the following distribution of breast composition in a series of 49,333 patients that underwent screening mammography: 10.5% almost entirely fat; 42.9% scattered fibroglandular densities; 38.7%

heterogeneously dense; and 7.5% extremely dense. The BCSC database reveals similar distributions of breast composition.[82]

- **Finding type**
  Prior published studies[83] have indicated that the distribution of mammographic findings in patients undergoing screening mammography is approximately as follows: 30-40% masses; 30-40% microcalcifications; 10-20% a combination of mass and microcalcifications, 10-20% architectural distortion or focal asymmmetry. The ACR BIRADS lists the following types of calcifications as intermediate concern (suspicious calcifications): amorphous or indistinct; coarse heterogeneous; and the following types of calcifications as higher probability of malignancy: fine pleomorphic; fine linear; fine-linear branching. Published series have shown wide variability of the types of microcalcifications seen on mammograms of patients subsequently shown to have breast cancer.[84]

- **Palpability**
  By definition, patients who undergo screening mammography are asymptomatic. It is well-known that palpable breast cancers are inherently different from non-palpable breast cancers.[85]

3. Mammography devices

Screen-film mammography

In screen-film mammography[86] (also known as analog mammography), an x-ray source exposes the breast to a beam of x-rays. The breast is positioned and stabilized by compression between the screen-film cassette and a compression paddle. A phosphor screen in a light-tight cassette absorbs a fraction of the x-rays that pass through the breast and are incident upon the screen. This fraction, typically 60% to 80%, is known as the quantum efficiency. The phosphor converts the absorbed x-rays to light. The light is coupled to a sheet of photographic film by direct contact of the screen and film within the cassette. The light emitted by the screen is recorded in the form of a latent photographic image on the film. The film is developed by chemical processing to produce a pattern of optical density on the film, which is then viewed by transillumination on a light box. The film serves as a recording, display and archive device. Screen-film mammography is still the most commonly used technique to obtain mammographic images.

Digital mammography

Digital mammography devices replace the screen-film detector with a digital (i.e., solid state) detector. Digital mammography devices that capture an image of the entire breast (in average size patients) are referred to as full-field digital mammography (FFDM) devices. Two basic designs of solid state detectors are currently in clinical use: [87]

i. Indirect flat-panel (the detector is composed of discrete sensing elements in the form of cesium iodide phosphor on large-area amorphous silicon active-matrix photodiode array); and
ii. Photo-stimulated storage phosphor (using a photostimulable phosphor plate with dual-side laser scan readout).

While the flat-panel system has physically discrete detector elements, the photo-stimulated storage phosphor system has a continuous detector that, when it is scanned by a laser beam for readout, the size of the beam and the distance it sweeps between sample measurements define the physical size of the "effective" detector elements. Digital detectors efficiently absorb x-rays, produce and process an electronic signal, and store the results in computer memory. The output image is saved as a 2-D matrix in which each element represents the x-ray transmission corresponding to a particular path through the breast.[88]

Manufacturers of digital mammography devices apply one or more processing algorithms to the digital data to "optimize" image display.[89] The digital data upon which such processing algorithms are applied is referred to as the "for processing data." The nature of these processing steps should be made clear to users.[90] The processed data is used to display (i.e., present) images to the radiologist. Once displayed, the image data can be further processed using such tools as histogram equalization, edge enhancement, and grayscale adjustments.

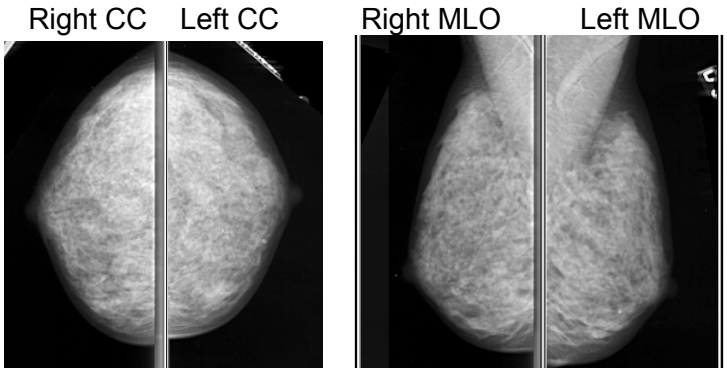
CAD processing of mammography images

Mammography CAD devices can operate on either screen-film or digital images. The first commercially available breast CAD device operated only on digitized screen-film mammograms and displayed marks on a small (approximately 5 inch diameter) monitor mounted below the film roller-panel. These devices could also output a paper printout with a small low-quality image with CAD marks. More recently, breast CAD devices operate on images from digital mammography devices and place marks directly on the digital images. It is important to note that both 'for processing' and 'for presentation' data can differ significantly between different FFDM systems and may affect CAD performance.

As discussed in Section II.A.1., the clinical impact of CAD may depend on the physical characteristics of the CAD mark.[91] Such physical characteristics would include the size and shape of the mark (e.g., rectangle, circle, asterisk, or actual segmentation from the CAD algorithm), the type of boundary (e.g., solid versus dashed), the color of the mark and the proximity of the mark to the potential abnormality. Visual search behaviors may change with CAD, potentially affecting workflow as well as the detection of other unmarked findings.[92] The number of false positive marks on a displayed image can also affect performance, with larger numbers of false positive prompts reducing true-positive detection rates.[93] Whether CAD marks should be viewed after an initial interpretation has been made or as soon as the images are available could affect radiologist performance. Mammography CAD devices typically place different types of marks for masses and microcalcifications. How the use of CAD impacts the amount of time spent on each image/patient is uncertain.

Mammographic projections

All mammographic examinations comprise at least two standard projections or views of each breast: a craniocaudal (CC) view and a mediolateral oblique (MLO) view. These views are named in relation to the direction of the x-ray beam during exposure. For a CC view, the x-ray beam passes through the superior aspect (i.e., cranial aspect) of the breast, while in compression, to the inferior aspect (i.e., caudal aspect) of the breast. For a MLO view, the x-ray source beam passes through the medial aspect of the breast in an oblique orientation to the lateral aspect of the breast. By convention, mammographic images are displayed such that the CC and MLO views are side-by-side as if the patient is facing the reader. Examples of CC and MLO views are reproduced below[94]:



4. Mammography interpretation

Radiologist task

Whether using screen-film or digital mammography systems, breast cancer is detected on the basis of four types of findings on the mammogram[95]:

- characteristic morphology of a tumor or mass
- the shape and spatial configuration of microcalcifications
- distortion of the normal architecture of the breast tissue
- asymmetry between views of the left and right breast

The primary goal of mammography interpretation is to accurately detect and describe these findings if they are present. At the same time, it is important not to falsely identify these findings if they are not actually present in the breast.

Mammography is unique among imaging tests because it must be performed (and even interpreted) in accordance with the Mammography Quality Standards Act (MQSA) regulations.[96] The MQSA establishes "minimum national quality standards for mammography facilities to ensure safe, reliable, and accurate mammography."[97] The quality standards not only regulate all aspects of personnel, equipment, quality assurance and exam performance, but also "medical records and mammography reports."[98] For example, the well-known final assessment categories of the ACR Breast Imaging Reporting and Data System (BI-RADS™ Atlas),[99] a collaborative effort involving many different organizations including FDA, mirror the MQSA requirements.[100] However, MQSA defines mammography as "radiography of the breast," but, specifically excludes "radiography of the breast performed during invasive interventions for localization or biopsy procedures." MQSA regulations do not apply to mammography CAD devices.

There are two types of mammographic examinations: screening and diagnostic. Screening mammography is "a radiological examination to detect unsuspected breast cancer in asymptomatic women … and should ordinarily be limited to craniocaudal (CC) and mediolateral oblique (MLO) views of each breast."[101] Screening mammograms are almost always performed without direct physician monitoring. On the other hand, diagnostic mammography "is a radiographic examination performed to provide additional information about patients who have signs and/or symptoms of breast disease, radiographic findings of concern, or in situations

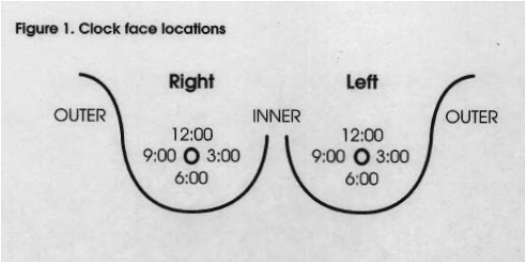
where direct supervision of the imaging is deemed appropriate by the interpreting physician. A diagnostic mammogram is performed under the direct supervision of a qualified physician in mammography and may include mediolateral oblique (MLO), craniocaudal (CC), and/or additional views."[102] Mammographic compression and magnification views are often used to evaluate microcalcifications, asymmetric densities, masses, or areas of architectural distortion.[103]

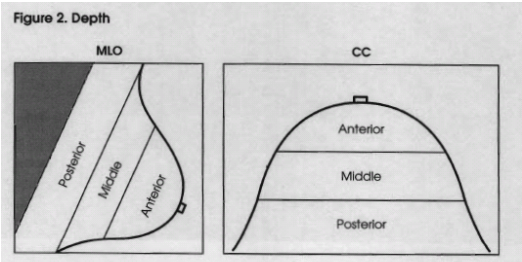
Whether performing screening or diagnostic mammography, the CC and MLO projections in each breast are considered complementary and necessary for proper interpretation. Mammography studies are almost always interpreted by examination of the CC views from each breast in a side-by-side manner (and likewise for the MLO views). If a potential abnormality is identified on a single view (either the CC or MLO), then the corresponding region on the complementary view is examined to confirm the three-dimensionality of the finding. In fact, the ACR BI-RADS defines a mass as "a space-occupying lesion seen in <u>two</u> different projections. If a potential mass is seen in only a single projection it should be called an asymmetry."[104] In the Guidance Chapter,[105] the ACR BI-RADS further clarifies that a mass is "a three-dimensional structure demonstrating <u>convex outward borders</u> … whereas asymmetries are planar, lack convex borders, usually contain interspersed fat and lack the conspicuity of a three-dimensional mass." The ACR BI-RADS employs "distribution modifiers" to describe calcifications and application of the modifiers requires examination of both the CC and MLO views. For example, the distribution modifier "grouped or clustered" calcifications is used when "at least five calcifications occupy a small volume (< 1.0 cc) of tissue."[106] Furthermore, side-by-side comparison should always be made between current and prior mammograms (when available).

<u>Mammography reporting</u>

All mammography examinations in the United States (screening and diagnostic) are interpreted and reported using the ACR BI-RADS Atlas. "The lexicon of mammography terms and the reporting format are meant to standardize the language used in mammography reports. In particular, the consistent use of the assessment categories will help clinicians understand disposition of their patients based on mammographic imaging and aid in auditing mammography practice. The illustrated BI-RADS® Fourth edition is designed for everyday practice and should make it possible to issue meaningful unambiguous mammography reports."[107]

Reporting the precise location of mammographic finding is critical for any follow-up examination and biopsy. "A significant lesion must always be triangulated so that its three dimensional location within the breast is known. … The location of the lesion should be described using clinical orientation extrapolated from film location. The breast is viewed as the face of a clock with the patient facing the observer. …. Depth divides the breast into anterior, middle and posterior thirds" as illustrated below.





Reporting the diagnostic assessment of findings is also critical. While each finding can be separately assessed, the final diagnostic assessment for the patient is based on the most worrisome finding(s) present.[108] The ACR developed and standardized the final diagnostic assessment into 7 BI-RADS assessment categories, as follows:

- Category 0 (Need Additional Imaging Evaluation and/or Prior Mammograms For Comparison)
- Category 1 (Negative)
- Category 2 (Benign Finding(s))
- Category 3 (Probably Benign Finding—Initial Short-Interval Follow-Up Suggested. A finding placed in this category should have less than a 2% risk of malignancy)
- Category 4 (Suspicious Abnormality—Biopsy Should Be Considered. Reserved for findings that do not have the classic appearance of malignancy but have a wide range of probability of malignancy that is greater than those in Category 3)
    - *Category 4A*: may be used for a finding needing intervention but with a *low suspicion for malignancy*. A malignant pathology report not expected and a 6-month or routine follow-up after a benign biopsy or cytology is appropriate.
    - *Category 4B:* includes lesions with an *intermediate suspicion of malignancy*. Findings in this category warrant close radiologic and pathologic correlation.
    - *Category 4C:* includes findings of *moderate concern, but not classic* (as in Category 5) for malignancy. Examples of findings placed in this category are an ill-defined, irregular solid mass or a new cluster of fine pleomorphic calcifications. A malignant result in this category is expected.
- Category 5 (Highly Suggestive of Malignancy—Appropriate Action Should Be Taken. These lesions have a high probability (≥95%) of being Cancer)
- Category 6 (Known Biopsy – Proven Malignancy—Appropriate Action Should Be Taken)

5. Performance measures of mammography

The Breast Cancer Surveillance Consortium (BCSC) is a collaborative network of seven mammography registries with linkages to tumor and/or pathology registries and contains information on approximately 6 million mammographic examinations (both screening and diagnostic), 2 million women, and 75,000 breast cancer cases.[109] Measures of screening mammography are based on 4,032,556 screening mammography examinations performed between 1996 and 2005 at 152 mammography facilities by 803 interpreting radiologists. Measures of diagnostic mammography performance are based on 714,984 diagnostic mammography examinations performed between 1996 and 2005 at 153 mammography facilities by 741 interpreting radiologists. The BCSC is funded and coordinated by the Applied Research Program of the National Cancer Institute (NCI) Division of Cancer Control and Population Sciences. BCSC data has been used to establish performance benchmarks for both screening and diagnostic mammography.[110] In addition to the BCSC data, there is extensive published literature on mammography performance measures.

There is great variability in radiologist interpretation of mammography. The sensitivity of screening mammography ranges from approximately 60% to 100% and the specificity ranges

from 35% to 98%.[111] The BCSC database shows an average sensitivity of 78.7% and an average specificity of 89.5% (where sensitivity was defined as the percentage of cancers that were BI-RADS category 0, 4, 5, or, 3 with a recommendation for immediate work-up; and specificity was defined as the percentage of non-cancers that were BI-RADS category 1, 2, or 3 without a recommendation for immediate work-up).[112] Using the definition of a missed cancer as one that becomes clinically evident within a year of a negative mammogram, screening mammography is about 70% to 75% sensitive in current clinical practice and about 10% of cancers are mammographically occult even after they are palpable.[113] Mammographic sensitivity is lowest in patients with dense breasts and for smaller cancers[114] and generally decreases as the recall rate (and false positive rate) decrease.[115] Sensitivity and false positives are also strongly dependent on observation times.[116]

The positive predictive value (PPV) of screening mammography ranges from approximately 2% to 22% for mammograms with any finding requiring further evaluation and from 12% to 78% for abnormal results requiring biopsy.[117] The BCSC database showed a PPV of 4.8% for an abnormal interpretation (BI-RADS 3— only when immediate work-up is recommended; or BI-RADS 0, 4, or 5; or when additional imaging was performed on the same day as the screening mammographic examination); a PPV of 25% when biopsy was recommended (based on the assignment of BI-RADS category 4 or 5 at the final assessment); and a PPV of 32.6% when biopsy was performed (defined as assignment of BI-RADS category 4 or 5 at the final assessment and availability of biopsy results).

Radiologist sensitivity is lower in patients with dense breasts and for smaller cancers. Sensitivity generally decreases as the recall rate (and false positive rate) decreases.[118] Sensitivity and false positives are strongly dependent on observation times.[119] In the study by Nodine et al,[120] mammographers detected true-positives four times faster than false-positives during the first 25 seconds of decision time. After 25 seconds, the false-positive rate became half of the true positive rate and continued to increase relative to the true-positive rate.

6. Radiologist errors on mammography examinations

Mammographic assessment is a difficult and complex task that essentially assesses the need for follow-up or biopsy.[121] When interpreting screening mammograms, the task of the radiologist is to detect and analyze findings rapidly and only spend significant time analyzing difficult findings, if any. Given that 90% to 95% of screening mammograms are either entirely normal or contain clearly benign findings, the task of the radiologist is to find the so-called “needle in a haystack.” A practicing radiologist may therefore perform detection and analysis very rapidly (i.e., both tasks performed almost simultaneously) in a clinical setting.[122]

On average approximately 20% of patients with breast cancer will have the cancer “missed” by mammography examination. However, some breast cancers are simply not visible (i.e., “invisible”) on mammography even under ideal conditions. For example, Durfee et al[123] found that 38/298 (12.8%) consecutive palpable breast cancers (i.e., known location at time of mammography) were not seen on mammography. When breast cancers are in fact “visible” on mammography, they can still be missed by radiologists due to cognitive, perceptual, technical or administrative errors (or some combination thereof). Ignoring technical and administrative type errors, unless one tracks the performance of the radiologist at the time of interpretation, simply reviewing radiology reports will not necessarily allow distinction between a true perceptual error (i.e., an error of detection) and a true cognitive error (i.e., an error of assessment). In other

words, a finding may be detected, analyzed and simply dismissed without mention in the report. Such an error could be misconstrued as an error of detection.

Prior publications have examined so-called "missed" breast cancers but have typically not distinguished between perceptual and cognitive errors as the cause. "Missed" breast cancers are typically identified by obtaining "prior" mammograms from patients with newly diagnosed cancer (i.e., patients with interval cancers) and determining if the cancers are "visible in retrospect."[124] As noted by Harvey et al.,[125] even though non-palpable breast cancers may be evident in retrospect on previous mammograms, many appear as an asymmetric density and may therefore not necessarily be true radiologic errors and that retrospective reviews do not reflect the everyday practice of screening mammography. Bird et al.[126] identified 320 cancers diagnosed in their screening population over a 5 year period and characterized 77 of the 320 cancers (24%) as "missed." Of the 77 missed cancers, 33 were listed as "overlooked." However, the other listed reasons for missed cancers make it clear that "overlooked" meant that the finding was not described in the prior radiology report.

Prospective visual analysis of cancers missed at screening mammography has revealed that cognitive errors (compared to perceptual errors) are more likely to account for missing a cancer.[127] That is, most cancers missed at screening mammography may in fact draw the radiologist's attention and be dismissed (with or without formal description in the radiology report). Such "missed" cancers would in fact not constitute errors of detection, but are rather the result of errors of analysis. For example, applying visual analysis (with tracking of eye movements) to both current and prior mammograms, Mello-Thomas et al.[128] found that "most unreported malignant lesions attracted some amount of visual attention, but it was in the processing of the information … that most errors occurred." A clinical study by Blanks et al.[129] found similar results. In any event, it is well-known that small masses (particularly in patients with dense breasts) likely account for the vast majority of "missed cancers" that are missed as a result of an error of detection.[130,131,132]

7. Origin of mammography CAD devices

Double reading of screening mammograms (i.e., reading by two radiologists) has been advocated as a way to increase radiologist sensitivity. Clinical studies have shown that double reading of mammograms improves radiologist detection by 5% to 15%[133] but typically with an associated increase in recall rate.[134] Double reading may be performed in various ways, including independent reading by two readers, one reader providing the primary interpretation and the second an overview, ''expert'' review by a second reviewer, consensus review by a team, and other variations.[135] Independent double reading is more frequently associated with a loss of specificity while consensus reading or expert double reading seems to improve sensitivity while maintaining specificity.[136] Double reading is not currently in widespread use in the U.S. due to practical and economic considerations.

As early as 1992, mammography CAD devices were proposed to potentially improve both detection and characterization.[137] CAD devices have subsequently been developed as a potential "replacement" for one of the double readers. If independent radiologist reading is performed first, CAD acts as a second reader with consensus expert opinion by the radiologist. If the radiologist reviews the CAD results before review of the mammogram, CAD acts first and the radiologist acts as a second, expert reader. In either case, the radiologist determines which

findings marked by CAD require further investigation with additional imaging or biopsy. However, the primary purpose of current CAD systems is to identify areas overlooked by the radiologist (i.e., detection errors) during routine interpretation.

Errors of observation include failure to observe more obvious (or routine) abnormalities as well as failure to observe more subtle abnormalities (i.e., there is a spectrum of the difficulty of detection of abnormalities that radiologists overlook). CAD devices could be designed to reduce one or both of these types of errors (i.e., primarily a portion of the spectrum or the entire spectrum). Such differences in design would also affect the rate (and type) of false positive marks. Such design differences (and user knowledge of the design) could significantly impact user interpretation. Furthermore, the sensitivity of CAD devices is well-known to be significantly higher for microcalcifications (>~90%) versus masses (or other lesions).

B. Mammography CAD devices approved by FDA

1. Currently approved devices

Four mammography CAD systems have been approved through premarket approval application (PMA) (approval orders in 1998, two in 2002, and 2004, respectively). The 'first-of-the-kind' device was the subject of a Radiological Devices Advisory Panel Meeting held on May 11, 1998.[138] All of these devices were first approved for use with digitized versions of screen film mammograms (i.e., not full-field digital mammography, FFDM, devices) obtained for screening purposes. Further PMA supplements were approved over time to expand the use of CAD devices to operate on digitized diagnostic screen film mammograms as well as mammograms obtained on FFDM devices. Supplements have also been approved for "updates" (i.e., new software versions) of the CAD algorithms.

The labeling of all approved mammography CAD devices contains Indications for Use Statements similar to the following:

"… intended to identify and mark regions of interest on routine screening and diagnostic mammograms to bring them to the attention of the radiologist after the initial reading has been completed. Thus, the software assists the radiologist in minimizing observational oversights by identifying areas on the original mammogram that may warrant a second review."

That is, all currently approved CAD devices are labeled for use as "second readers" (meaning that the radiologist is expected to perform a complete evaluation of a mammographic study, then turn on the CAD device, and then additionally assess any regions on the mammogram "marked" by the device).

Note that for use in diagnostic mammography, CAD devices are only approved for symptomatic patients with standard mammographic views (CC and MLO).

2. Data and information provided for original approval of mammography CAD devices

All PMAs are required to contain data and information that constitute valid scientific evidence and provide reasonable assurance that the device is safe and effective for its intended use. The evidence required may vary according to the characteristics of the device, its conditions of use, the existence and adequacy of warnings and other restrictions, and the extent of experience with its use. At the time of approval of the first mammography CAD device in 1998, there was little if any experience with its use by radiologists in clinical practice.

The data that served as the basis for approval for currently approved mammography CAD devices included four components: three 'standalone' studies (see items a, b, and c below) and one reader study that assessed recall rate in patients without cancer but in which device sensitivity was not measured (see item d).

a. Standalone performance on "missed" cancer cases was used to determine the ability of the CAD device to mark the cancers. The "missed cancers" were defined as cancers that were retrospectively "visible" and "actionable" (where "actionable" meant not BI-RADS 1 or 2) on the prior mammograms of patients with newly diagnosed cancer. The visibility and actionability of the "missed cancers" was determined by a panel of radiologists who knew a cancer had been missed on the prior mammograms. The ability of the CAD devices to mark the cancer on the prior mammograms was then measured in a standalone manner. This was considered a surrogate for the ability of the device to detect "difficult" findings and was used to estimate the <u>potential</u> reduction of errors of detection for radiologists using the CAD device in practice.

b. Standalone performance on cancers detected at screening mammography was used to determine the ability of CAD devices to also mark cancers that radiologists typically detect on screening mammograms. This was considered a surrogate for the ability of the device to detect more obvious and intermediate level of difficulty lesions.

c. Standalone performance on normal screening mammograms was used to determine the rate of false positive CAD marks on cases that did not have a known cancer.

d. Routine screening exams (with or without enrichment with some cancers) were used to determine the potential increase in recall rate resulting from use of the CAD devices. Cases were read by multiple radiologists, without and with a CAD device and the recall rates were compared.

Since 2001, there has been published literature discussing the potential limitations of these standalone performance studies.[139] For example, the published literature has noted that separation of (1) the cohort of patients used to measures the potential reduction in missed cancers from (2) the cohort of patients used to measure the effect on recall rate, produces conditions that do not provide a statistically interpretable result.[140] Under such testing conditions, there is no control (or accounting) of the degree of case difficulty or the range of reader skill and level of reader aggressiveness across the two measurements. In Section III.D., below, we ask the Panel's input on what testing should be recommended to support approval of future mammography CAD devices.

C. Published literature on approved devices

There is now a large body of published literature on the subject of clinical performance testing of mammography CAD devices where radiologist performance is measured both without and with use of the CAD device. These studies employed two different general designs: retrospective clinical performance testing and prospective clinical performance testing. The retrospective clinical studies use radiologist interpretations that are not part of an actual clinical practice. The prospective clinical studies use radiologist interpretations that are part of an actual clinical practice. Furthermore, the prospective clinical studies include *sequential* or *historical control* designs. The sequential design presents the radiologist with an image without CAD information, requires interpretation, and then presents the same image with CAD markings and allows the radiologist to modify the assessment. The historical control design compares radiologist performance over a period of time without CAD devices to radiologist performance over a period of time after CAD introduction. The historical control design will obviously have entirely different patients over the two time periods.

A recent (November 2007) comprehensive review article by Helvie includes a summary of previously published studies examining clinical performance testing of mammography CAD performance. We add a few minor modifications and additions from the recent literature to the following key points of Helvie:

- Standalone testing of mammography CAD devices reveals high sensitivity to mark microcalcifications but lower sensitivity to mark masses (or other abnormalities).
- Standalone testing of CAD devices reveals a FP mark rate of between 2 and 4 marks per patient (i.e., 0.5 to 1.0 marks per view of a 4-view mammogram).
- Studies evaluating clinical performance have not demonstrated statistically significant improvement of radiologist sensitivity for detection of invasive cancers that manifest as masses or microcalcifications.
- Some of the studies evaluating clinical performance show a trend (not statistically significant) toward CAD improving radiologist detection of DCIS.
- Studies evaluating clinical performance show an increase of the recall rate when using CAD devices; in some studies, these increases are statistically significant.

Note that because of the low prevalence of breast cancer cases in a screening population, statistical significance may be difficult to achieve without a large study.

Copies of Helvie's review and the original publications are provided to the Panel. Brief summaries of the articles are as follows.

Retrospective clinical performance testing

- In 1998, Thurfjell et al.[141], reviewed the effect of CAD on an enriched set of 120 mammograms (74 cancers and 46 normals) read by 3 radiologists (expert, a screening radiologist, and a clinical radiologist). The standalone sensitivity for CAD was 50% compared with 59% for the expert. CAD resulted in no incremental detections for the expert, but did result in a 9% detection improvement for the least experienced radiologist.

- In 2003, Ciatto et al.[142], used CAD in an enriched data set of 150 cases (including 17 cancers). Studies were read by 10 radiologists without and with CAD. There was a 4.2% increase (p=0.31) in sensitivity noted. There was a 44% (p<0.003) increase in the recall rate from 7.9 to 11.4%.

- In 2004, Taylor et al.,[143] applied a CAD system to a test set of 180 cases (including 60 cancers; 20 false negative interval cases and 40 screen-detected cancers) to 30 radiologists. No significant difference was found for readers' sensitivity or specificity between the prompted and unprompted conditions. No significant difference was found between the sensitivity and specificity of the different groups of readers. No difference in impact of prompts was found for well or poorly performing readers.

- Alberdi et al.[144] attempted to determine the effect of FN CAD assessments. An experimental data set of 60 cases (including 30 cancers) was assessed by 39 physicians. The researchers modified the FP marks and then presented the cases to radiologists with CAD annotations or without CAD in a test situation. Average reader sensitivity with CAD decreased significantly compared with reading without CAD (61% versus 73%). Specificity was without significant change (90% with CAD and 86% without CAD). Most of the decline in sensitivity was for cases that were FN by CAD.

- Destounis et al.[145] attempted to determine whether CAD provides incremental information for cases that have already undergone double reading. In this situation, CAD is acting as a ‘‘triple’’ reader. CAD was applied retrospectively to 52 screening cancer cases not originally detected by clinical double reading. In this experiment, 71% of these screening mammograms were retrospectively marked correctly by CAD, suggesting the potential for CAD to reduce FN cases, even for those cases assessed by double reading.

Prospective reader studies

- Freer and Ulissey[146] reported the first sequential clinical CAD study of a community practice. Two readers examined each screening mammogram sequentially without and with CAD. They noted a non-significant 19.5% improvement in mammographic cancer detection in a study of 12,860 patients (radiologists without CAD detected 41 cancers; radiologists with CAD detected 49 cancers). Of the incremental cancers detected by CAD, 87% were microcalcifications; 75% of incremental detections (six of eight) were DCIS. There was a 5% invasive cancer detection improvement. Specificity was noted to decline with CAD. Recall rates increased by 18.5%, and short-term follow-up for BI-RADS Category 3 probably benign cases increased by 38%. CAD failed to mark 33% of invasive cancers manifest as masses but found by the radiologist. Overall, CAD marked 82% of cancers.

- Gur et al.[147] reported a large clinical trial involving 24 radiologists and 115,571 examinations. They found no significant difference in the malignancy detection rate or recall rate in a study using a historical control model following institution of a commercial CAD program at a large academic practice. A 1.7% increase in cancer detection rate was observed with CAD (3.49/1000 without CAD versus 3.55/1000 with CAD; p < 0.68). The recall rate (11.4%) increased by 0.1%. Although subset analyses noted sensitivity improvement for low-volume radiologists, high-volume readers had a non-significant decrease in their sensitivity with CAD.

- Birdwell et al.[148] reported on 8682 consecutive cases using the sequential methodology during a 19-month period at Stanford University. There were 29 cancer cases: 27 detected by the radiologist without CAD and 2 additional cancers detected with CAD. The investigators found a non-significant 7% increase in cancer detection and an associated 8% increase in recall rate. The overall cancer detection rate was 3.3/1000. CAD detected 79.3% of cancers. Both incremental cancers prompted by CAD were invasive masses.

- Cupples et al.[149] reported a historical control study involving 7872 patients examined before institution of a CAD program and 19,402 patients following the introduction of a commercial CAD program in a community practice. A non-significant increase in the overall cancer detection rate from 3.7/1000 to 4.3/1000 with CAD (16%) was found (the authors also noted a non-significant increase in the detection rate of invasive cancers of 1.0 cm or less in size). There was an increase in the recall rate from 7.7% to 8.3% and an increase in the biopsy rate from 1.4% to 1.5%. Forty-seven percent of the CAD cases were examined by a reader not present in the pre-CAD arm of the study; the readers present during both study periods had sensitivity improvement.

- Morton et al.[150] reported a sequential reader study using a commercial CAD system. They found a non-significant 7.6% increase in cancer detection (from 4.92/ 1000 to 5.29/1000). The recall rated increased 9.5%, from 9.84% to 10.77% with CAD. PPV remained stable at 41%. In this trial, the commercial CAD system marked 76% of cancers, which included 65% of masses and 100% of calcification cases. Five of eight (63%) cancers not detected by routine care but detected by CAD were microcalcifications, and all were low- to intermediate-grade DCIS. There was a 2.8% increase in the detection of invasive cancers (3/105).

- Dean et al.[151] reported a single institution experience for CAD in screening and diagnostic cases in a private practice setting. This was a sequential reader study design and all readings were performed by a single reader. Thirty-five cancers were detected in 5,631 screening patients, 33 of which were nonpalpable. In the 3,803 diagnostic mammograms, 63 cancers were detected, of which 36 were palpable and 27 were nonpalpable. A non-significant 13% improvement in cancer detection was noted for 5631 screening cases. The recall rate increased 26%, from 6.2% to 7.8%. CAD marked 89% of all screening cancer cases and 83% of combined screening and diagnostic cases. CAD detected 64% of invasive lobular cancers and 83% of invasive ductal cancers. Overall, 30% of CAD incremental detections were for DCIS. DCIS was the only type of cancer for which CAD noted more cases than the radiologist.

- Ko et al.[152] reported a sequential CAD experience involving 5016 screening cases (45 cancers). A non-significant 4.7% increase in cancer detection was matched by a 14.7% increase in the recall rate. CAD resulted in a 2% increase in incremental invasive cancer clinical detection (50% of CAD incremental cases were DCIS). Like other studies, CAD performance for malignant calcification cases was greater than for masses (100% versus 67%). In this study, CAD produced one mark per image.

Recent studies

- Fenton et al.[153] (April 2007) determined the association between the use of computer-aided detection and the performance of screening mammography from 1998 through 2002 at 43 facilities in three states. They had complete data for 222,135 women (a total of 429,345 mammograms), including 2351 women who received a diagnosis of breast cancer within 1

year after screening. They calculated the specificity, sensitivity, and positive predictive value of screening mammography with and without computer-aided detection, as well as the rates of biopsy and breast-cancer detection and the overall accuracy, measured as the area under the ROC curve. Seven facilities (16%) implemented computer-aided detection during the study period. Diagnostic specificity decreased from 90.2% before implementation to 87.2% after implementation ($p < 0.001$), the positive predictive value decreased from 4.1% to 3.2% ($p = 0.01$), and the rate of biopsy increased by 19.7% ($p < 0.001$). The increase in sensitivity from 80.4% before implementation of computer-aided detection to 84.0% after implementation was not significant ($p = 0.32$). The change in the cancer-detection rate (including invasive breast cancers and ductal carcinomas in situ) was not significant (4.15 cases per 1000 screening mammograms before implementation and 4.20 cases after implementation, $p = 0.90$). Analyses of data from all 43 facilities showed that the use of computer-aided detection was associated with significantly lower overall accuracy than was non-use (area under the ROC curve, 0.871 vs. 0.919; $p = 0.005$).

- Georgian-Smith et al.[154] (November 2007) compared a human second reader with a CAD using a retrospective review of 6,381 consecutive screening mammograms that had been interpreted by a primary reader. The primary reader then reinterpreted the studies using a CAD device ("CAD reader"). A second human reader (blinded to the CAD results but knowledgeable of the primary reader's findings) also reviewed the studies, looking for abnormalities not seen by the first reader. Two cancers were called back by the second human reader that were not called back by the CAD reader; however, the CAD system had marked the findings, but they were dismissed by the primary reader. The difference between the CAD reader and second human reader was not statistically significant. The CAD and human second readers increased the recall rates 6.4% and 7.2% ($p = 0.70$), respectively, and the biopsy rates 10% and 14.7%. The positive predictive value was 0% (0/3) for the CAD reader and was 40% (2/5) for the human second reader. The relative increases in the cancer detection rate compared with the primary reader's detection rate were 0% for the CAD reader and 15.4% (2/13) for the human second reader ($p = 0.50$).

- Ellis et al.[155] (October 2007) retrospectively compared two commercially available CAD systems for detecting invasive breast cancers manifesting as non-calcified masses smaller than 16 mm. The respective per study sensitivity, per image (i.e., per view) sensitivity, per study specificity, and mass false positive marker rates were 81.8%, 64.7%, 39.2%, and 1.08 for one of the CAD devices, and 60.9%, 42.6%, 31.4%, and 1.41 for the second CAD device. The overall per study and per image sensitivities were significantly better for one of the devices (McNemar test, all $P < 0.001$), with a non-significant higher per study specificity and lower mass false marker rate on normal studies. CAD results demonstrated at least a 20% variation between BI-RADS categories 4a and 5 for per study and per image sensitivity.

- Taplin et al.[156] (December 2006) retrospectively created a stratified random sample of screening mammograms weighted with difficult cases split evenly among women with fatty breast tissue and those with dense breast tissue: 114 patients were cancer-free, 114 had cancer 1 year after screening, and 113 had cancer 13–24 months after screening. The objective of the study was to test whether the use of CAD improves sensitivity at no cost to specificity for the detection of breast cancer and enables more accurate assessment of fatty breast tissue compared with dense breast tissue. In test settings 6 months apart, 19 community radiologists interpreted 341 bilateral screening mammograms with and without CAD. CAD assistance did not affect overall sensitivity (cancer by 1 year: 63.2% without CAD and 62.0% with CAD; cancer in 13–24 months: 33.5% without CAD and 32.3% with

CAD). Overall specificity was 72% without and 75% with CAD ($p < 0.02$). CAD had a greater effect on both specificity ($p < 0.02$) and sensitivity ($p < 0.03$) among radiologists who interpret more than 50 mammograms per week. The results were the same for fatty breast tissue and dense breast tissue. The authors concluded that CAD increased interpretive specificity but did not affect sensitivity because visible non-calcified lesions that went unmarked by CAD were less likely to be assessed as abnormal by radiologists. Breast density did not affect CAD performance.

Gromet[157] compares the efficacy of single reading with computer-aided detection to double reading and also to first reader (without CAD) in a study with 232,221 screening mammograms. This recent publication is provided to panel members for review, although its content was not reviewed or summarized in this document.

D. Clinical testing issues specific to mammography CAD

1. Ground truth

Evaluation of both standalone performance testing and clinical performance testing of mammography CAD devices relies on determining whether or not disease (or an image finding) is present within an individual patient, and the location and extent of that disease or finding. In the case of mammography CAD devices, the disease is breast cancer and the imaging findings are masses and microcalcifications.

Ground truth identifies (a) whether or not a patient has cancer, (b) whether or not the cancer is revealed on the mammogram as a mass (or microcalcifications, architectural distortion or asymmetry), and (c) the precise location and extent of the findings. Ground truth for breast disease is typically determined by biopsy or surgery for malignant (or benign) findings and by a negative one-year follow-up for normal patient (or benign findings)

2. Standalone performance testing

Standalone performance measures how well the mammography CAD device marks regions of known pathologic/mammographic findings (and how well the CAD device avoids marking regions without pathologic/mammographic findings) in the absence of radiologist interaction. To measure standalone performance, the true location of abnormalities should be determined through some well-described ground truth process (see Section III.D.1, above). The location of a CAD mark is compared to the true location of the pathologic/imaging finding and an assessment is made to determine if the CAD mark in fact sufficiently overlaps the true location of the finding (i.e., scoring).

As noted in Section III.A.4., above, breast cancer is detected on the basis of four types of findings on the mammogram: mass, microcalcifications, architectural distortion and asymmetries (i.e., asymmetry between the left and right breast). The ability to detect each of these findings may be affected by breast density and finding size. The pathologic evaluation will classify cancers as invasive or in situ (e.g., DCIS). Clinical, mammographic and pathologic information is also used to determine the Stage of the cancer (e.g., early or late disease at the time of diagnosis).

Standalone performance testing may be accomplished using a larger database than clinical performance testing. This may allow meaningful testing and reporting of standalone performance on clinically important subgroups (e.g., small masses in dense breasts), which helps determine user confidence in the particular types of marks (e.g., masses versus microcalcifications). One possibility for stratification might be to stratify results according to mammographic finding type (mass, microcalcifications, architectural distortion and asymmetries), size of mammographic findings, and/or breast composition.

Topic for discussion

M1. What substrata, if any, should be considered in standalone performance testing and labeling of mammography CAD devices?

Standalone performance can be reported on a per lesion, per view, per breast (2 views) or per patient (2 breasts with 2 views each) basis. Using a per lesion basis provides single definitions of TP, FP, TN, and FN. With a per breast or per patient basis, there can be various definitions of TP, FP, TN, and FN that will result in different standalone performance results (See Section II.C.5). For example, depending on the definitions used with a per breast basis, if a finding is revealed on both the CC and MLO views and the mammography CAD device marks the finding on only one view, this could result in a TP and a FN for per breast reporting or it could be counted as a TP if reported per breast.

Topic for discussion

M2. Please discuss the merits of per lesion, per view, per breast, and per patient endpoints for the standalone testing of mammography CAD.

Standalone performance depends upon determining whether a CAD mark actually 'overlaps' the true finding (scoring methodology). For example, a very large CAD mark can easily overlap or contain a small finding; a very small CAD mark can easily be contained by a very large finding

Topic for discussion

M3. Does the Panel have any recommendation on the marking or scoring methodology that should be used for reporting standalone performance for mammography findings?

3. Reader performance testing

Standalone performance testing indicates how well the device marks locations of interest. However, standalone performance does not indicate what locations will be considered suspicious by the physician using the CAD device. As discussed in Section II.C.6., there are several types of reader tests that are designed to determine the impact of a CAD device on physician performance, e.g. prospective studies, retrospective MRMC studies, retrospective Burhenne design studies.

Topic for discussion

M4. Are both device standalone performance and reader performance testing necessary to demonstrate that mammography CAD devices provide clinically significant results? Would one or the other suffice? Are there other types of studies that should be provided instead or additionally?

Test dataset

The prevalence of breast cancer cases in a screening population is approximately 0.6%[158] to 0.8%.[159] To limit the size of test datasets to reasonable numbers, evaluation of mammography CAD devices is often performed using a test dataset that is enriched with cases representative of the true population, but with a significantly greater percentage of patients with breast cancer (i.e., a population of patients with a prevalence of cancer much higher than that in a true screening population). These enriched dataset include non-cancer patients with normal and benign mammographic findings.

Ignoring prevalence, if the cancer and non-cancer patients in the enriched population should otherwise have clinical, mammographic and pathologic characteristics that are typically seen in a screening population (see Section III.A.2), such testing simulates a so-called "field test" where a field test is a real-time real-life assessment of a system in its most natural and representative habitat.[160]

An enriched database might seek to represent the following characteristics of the screening population:

- breast density: 40-50% of patients with heterogeneously dense or extremely dense breasts
- proportion of masses and microcalcifications: approximately evenly distributed with a sufficient number of additional patients with architectural distortion alone
- mass size: < 1.0 cm for the majority of the cases
- size and type of microcalcification clusters: small clusters of up to five microcalcifications for a third of the cases
- lesion size: mammographic and pathologic lesion size
- types of microcalcifications according to the American College of Radiology (ACR) BIRAD descriptors: punctuate, fine linear, round, etc
- focal asymmetries and architectural distortions: 10-20%
- normal, benign and malignant cases
- histologic types.

When enrichment is performed with only difficult cases, the testing is often referred as a stress test. In this situation, the test dataset is enriched with difficult cases that would be expected to challenge the readers and establish where the CAD is most beneficial. Accordingly, the patient dataset contains cancer cases that are more difficult to detect mammographically:[161] A stress test dataset might include:

- breast density: more patients with heterogeneously dense or extremely dense breasts
- proportion of masses and microcalcifications: more masses
- mass cases: more small masses
- non-cancer cases could included patients with benign findings that may be confused

with cancer (e.g., BI-RADS 4 and 5 cases where biopsy shows no cancer) as well as a number of normal patients to serve as distractors and reduce the prevalence of suspicious findings.

Topic for discussion

M5. Is an enriched reader study or a stress test an acceptable alternative to a field test? If so, are the following the appropriate clinical and mammographic characteristics of the screening population that should be considered when designing an enriched database or stress test? Are there other characteristics that should be considered?

- breast density: 40-50% of patients with heterogeneously dense or extremely dense breasts
- proportion of masses and microcalcifications: approximately evenly distributed with a sufficient number of additional patients with architectural distortion alone
- mass cases: non-palpable and a majority with size < 1.0 cm
- microcalcification cases: small clusters of up to five microcalcifications for a third of the cases
- type of microcalcification clusters: representation of types of microcalcifications according to the American College of Radiology (ACR) BI-RADS descriptors, e.g., punctuate, fine linear, round, etc.

Study endpoints

Topics for discussion

M6. Please discuss the merits and practicality of ROC analysis (full curve, area under the curve, partial area under the curve, sensitivity and specificity) as a primary endpoint(s). Are there alternatives? If so, please discuss their merits. What are clinically significant effect sizes?

M7. Please discuss the merits of per patient, per breast, per view, and per lesion endpoints for the mammography screening task. If per patient analysis is used, should the study endpoints also include an assessment of location accuracy?

M8. Should data be analyzed separately to measure the effectiveness of CAD devices for (a) detection of cancers that manifest as masses and (b) detection of cancers that manifest as microcalcifications?

M9. CAD use can impact reading time. Should reading time be assessed? If so, how?

Reader paradigms

Topic for discussion

M10. FDA does not specify indications for use, but reviews indications for use that are requested by Sponsors. What are the Panel's views regarding second reader versus concurrent reading using a CAD device? Specifically,

a. How are mammography CADs used clinically?
b. Are second reader and concurrent reading modes both clinically relevant options for use in practice? If not, which paradigm(s) are appropriate for mammography CAD devices?
c. Do you believe users understand that if a device is labeled as a second reader, the (i.e. the physician) should always read the radiological image completely before turning on the CAD?

4. Other issues for mammography CAD devices

Images obtained using screen film devices show little variability from device to device. However, screen film and FFDM have different spatial and contrast resolution and dynamic range. FFDM systems also vary in spatial and contrast resolution because of differences of solid state detectors, pixel sizes, and quantum and electronic noises. Furthermore, different FFDM manufacturers use different image processing algorithms, and CAD devices work with the raw FFDM and different image processing algorithms to identify abnormalities.[162],[163] Therefore, the effectiveness of CAD devices tested on screen film devices may differ from effectiveness on FFDM devices, and effectiveness may also vary on different FFDM devices.

Topic for discussion

M11. If a mammography CAD has been approved to operate with screen film or specific FFDM device(s), what data should be used to assess its performance with a different FFDM device? Are both device standalone performance and reader performance testing necessary? Would one or the other suffice? Are there other types of studies that should be provided instead or additionally?

Topic for discussion

M12. Based on the published studies of approved mammography CAD devices, is there anything else that should be considered regarding testing for future mammography CAD devices?

IV. Lung CAD devices

A. Introduction of lung CAD into the clinical practice

1. Lung imaging

Radiographic studies of the lungs are by far the most common type of imaging test in the United States. This is primarily attributable to the chest x-ray, the most common of all radiology tests.[164] Indications for chest x-rays in hospitalized patients, are vast and diverse[165]:

- monitoring stable patients
- respiratory failure (e.g., patients receiving mechanical ventilation)
- compromised respiratory function (e.g., patients with endotracheal tubes)
- central venous pressure catheter (CVP) insertion
- cardiopulmonary compromise with Swan-Ganz catheter insertion

Indications for emergency chest x-rays in non-hospitalized patients include:

- acute chest pain (e.g., suspected myocardial infarction)
- shortness of breath (e.g., suspected heart failure or pulmonary embolus)
- fever (e.g., suspected pneumonia)
- trauma

Indications for non-urgent chest x-rays in non-hospitalized patients include:

- screening for pulmonary metastases
- monitoring lung disease
- monitoring cancer therapy

Chest x-ray is a very useful but limited examination that provides poor soft tissue contrast and suffers from overlap of structures on a two-dimensional projection. Therefore, additional imaging tests are frequently necessary to clarify chest x-ray findings or to search for abnormalities not visible on the chest x-ray.

Computed Tomography (CT) scans provide improved soft tissue contrast and disease detection compared with chest x-rays. They are currently labeled for use to provide diagnostic information in symptomatic patients – they are not cleared or approved by FDA for any screening purpose. Chest CT may be a complementary examination to other imaging studies such as chest radiography or a standalone procedure.[166]The use of chest CT includes, but is not limited to[167]:

- Evaluation of abnormalities discovered on chest radiographs
- Evaluation of clinically suspected occult thoracic pathology
- Staging and follow-up of lung and other primary or secondary thoracic malignancies
- Evaluation for thoracic manifestations of known extrathoracic disease

- Evaluation of known or suspected thoracic vascular abnormalities (congenital or acquired)
- Evaluation and follow-up of pulmonary parenchymal and airway disease;
- Evaluation of trauma

For more precise evaluation of a variety of pulmonary diseases, the ACR recommends High-Resolution Computed Tomography (HRCT) of the lungs. The use of HRCT of the lungs includes, but is not limited to, the following[168]:

- Evaluation of diffuse pulmonary disease discovered on chest radiographs, conventional CT of the chest, or other CT examinations that include portions of the chest, including selection of the appropriate site for biopsy of diffuse lung disease
- Evaluation of the lungs in patients with clinically suspected pulmonary disorders with normal or equivocal chest radiographs
- Evaluation of suspected small airway disease
- Evaluation of suspected bronchiectasis
- Quantification of the extent of diffuse lung disease for purposes of evaluating effectiveness of treatment

2. Lung cancer and pulmonary nodules

More people die from lung cancer in the U.S. than any other type of cancer.[169] This is true for both men and women. In 2004 (the most recent year for which statistics are currently available), lung cancer accounted for more deaths than breast cancer, prostate cancer, and colon cancer combined. In 2004, 108,355 men and 87,897 women were diagnosed with lung cancer and 89,575 men and 68,431 women died from lung cancer.[170] There is some evidence that early detection of lung cancer may result in a more favorable prognosis.[171] This has led some to propose lung cancer screening in high-risk patients using chest computed tomography (CT) scans. This remains a topic of debate in the literature.[172]

The solitary pulmonary nodule is traditionally defined (on chest x-ray) as a relatively spherical opacity 3 cm or less in diameter surrounded by lung parenchyma.[173] However, according to the Fleischner Society, “the lung nodule has eluded all efforts at precise definition.”[174] The Fleischner Society on CT nomenclature described the pathologic definition of a nodule as: a “small, approximately spherical, circumscribed focus of abnormal tissue” and the radiologic definition as a “round opacity, at least moderately well marginated and no greater than 3 cm in maximum diameter.”[175]

Patients with a solitary pulmonary nodule (SPN) rarely have symptoms referable to the nodule and so detection is usually unexpected and in pursuit of other concerns.[176] The ever-expanding use of CT in medical imaging has led to an expansion of the definition of the SPN.[177] The more generic term to describe a “nodule,” is a “focal opacity.” A focal opacity includes abnormalities that are solid, semi-solid and non-solid (i.e., ground glass opacity). Based on chest x-rays, the incidence of SPN’s was traditionally placed in the range of 150,000 new cases per year in the United States.[178] However, this figure does not include all of the smaller nodules detected with CT and does not account for the fact that many patients have multiple “nodules” on chest CT scans. Therefore, inclusion of

the smaller CT-detected nodules would dramatically increase the incidence of pulmonary nodules, whether solitary or multiple. Although a precise definition of "small" has not been standardized, it is generally considered to be in the range of less than 1 cm.[179]

The primary concern in evaluating even smaller nodules is the ability to exclude malignancy and this may require CT-guided biopsy. Unfortunately, a solitary pulmonary nodule may be secondary to one of a long list of differential diagnoses:

- *Neoplastic (malignant or benign)*:
  - bronchogenic carcinoma
  - metastasis, lymphoma
  - carcinoid, hamartoma
  - connective tissue and neural tumors (fibroma, neurofibroma, blastoma, sarcoma)

- *Inflammatory (infectious)*:
  - granuloma (TB, histoplasmosis, coccidioidomycosis, blastomycosis, cryptococcosis, nocardiosis)
  - lung abscess
  - round pneumonia
  - hydatid cyst

- *Inflammatory (noninfectious)*:
  - rheumatoid arthritis
  - Wegener's granulomatosis
  - sarcoidosis
  - lipoid pneumonia

- *Congenital:*
  - arteriovenous malformation
  - sequestration
  - lung cyst

- *Miscellaneous*:
  - pulmonary infarct
  - round atelectasis
  - mucoid impaction
  - progressive massive fibrosis

3. Chest x-ray and CT imaging of pulmonary nodules

Small pulmonary nodules are often missed on chest x-rays because of poor lesion conspicuity caused by superimposition of hilar and mediastinal structures, blood vessels, clavicles, or ribs.[180] Poor viewing conditions and poor technical quality of the images are also reasons why small lesions can be missed. Published studies reveal that the proportion of missed lung cancers on chest x-rays varies between 25% and 90%.[181] When a nodule (or suspected nodule) is detected on chest x-ray, patients are typically referred to CT for further evaluation. However, the vast majority of nodules detected on CT are not in patients with prior chest x-rays revealing a nodule but are either incidental

or associated with known thoracic or extra-thoracic disease. Non-calcified lung nodules are quite frequent in smokers (or non-smokers with a prior significant smoking history) with a frequency ranging from 23-69%.[182]

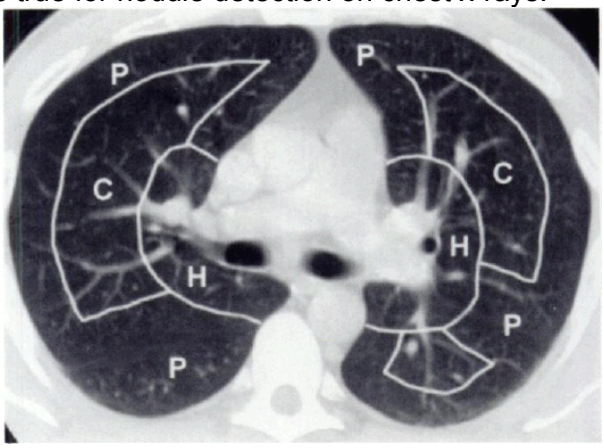
Whether first detected on chest x-ray or CT, the management of pulmonary nodules is problematic. Traditional medical practice based on chest x-ray imaging had been to regard all non-calcified pulmonary nodules as potentially malignant lesions that required close monitoring until shown to be stable over a period of at least 2 years.[183] This practice preceded the widespread use of CT because a substantial proportion of non-calcified nodules detected on chest x-ray were lung cancers. However, those nodules that developed into cancer were almost all larger than 5 mm in diameter (most were 1–3-cm in size).[184] When using CT findings to manage patients with a pulmonary nodule, size and access for biopsy are key considerations. In addition, any calcification in a small nodule favors a benign cause and central, laminar, or dense diffuse patterns of calcification are reliable evidence of benignancy[185] and fat content suggests a hamartoma, lipoid granuloma or lipoma.[186] For CT detected nodules less than 8 mm in size, the most recent management recommendations for an indeterminate nodule in persons 35 years of age or older include the following:[187]

Nodule Size	Low Risk Patient	High Risk Patient
≤ 4 mm	No follow-up needed	Follow-up CT at 12 mo; if unchanged, no further follow-up
5 – 6 mm	Follow-up CT at 12 mo; if unchanged, no further follow-up	Initial follow-up CT at 6–12 mo then at 18–24 mo if no change
7-8 mm	Initial follow-up CT at 6–12 mo then at 18–24 mo if no change	Initial follow-up CT at 3–6 mo then at 9–12 and 24 mo if no change
> 8 mm	Follow-up CT at around 3, 9, and 24 mo, dynamic contrast enhanced CT, PET, and/or biopsy	Same as for low-risk patient

For nodules larger than 8 mm, unless contraindicated, CT-guided biopsy is typically performed.

When interpreting chest CT studies, many factors affect radiologist sensitivity for nodule detection including size, location and proximity to blood vessels.[188] CT sensitivity for the detection of nodules 10 mm or larger is nearly 100%.[189] In a study by Rusinek et al.[190] (using simulated nodules on chest CT images) the location of each nodule within the lung was classified as being peripheral (all lung within 2 cm of the pleural surfaces), perihilar (all lung parenchyma within 2 cm of the hilum), or central (all remaining panenchyma), see figure below from Rusinek et al. They calculated detection sensitivity as a function of size, location and proximity to blood vessels at both low dose (20 mAs) and conventional dose (200 mAs). With regard to size: for both low and conventional dose, detection sensitivity was significantly greater for 7 mm versus 3 mm nodules. With regard to location: for both low and conventional dose, detection sensitivity was greatest for peripheral nodules and lowest for perihilar nodules. With regard to proximity to blood vessels: for both low and conventional dose, detection sensitivity was greater for nodules that were adjacent to vessels and lower for nodules that were separate from vessels. With regard to dose: sensitivity is significantly reduced for peripheral nodules ($p = .019$) and nodules separated from blood vessels ($p = .044$) on low dose CT scans. These results were similar to those published by Naidich et al[191] (but where dose was not a variable).

Therefore, radiologist detection sensitivity on CT has the greatest room for improvement when dealing with: small nodules, perihilar nodules and nodules adjacent to blood vessels. The same is true for nodule detection on chest x-rays.[192]



4. CAD devices for chest x-rays and chest CT

Both chest x-ray and CT lung CAD systems have been introduced into clinical practice as a second reader (after the physician has performed an initial interpretation of the patient chest data) to aid/assist the physician's identification/detection of solitary/solid pulmonary nodules. Below are example excerpts from currently approved/cleared Indications for Use statements:

- **Chest x-ray CAD**: intended to identify and mark regions of interest (ROI) on digitized frontal chest radiographs. It identifies features associated with solitary pulmonary nodules from 9 to 30 mm in size, which could represent early-stage lung cancer.

- **CT lung CAD**: designed to assist radiologists in the detection of solid pulmonary nodules during review of multi-detector CT (MDCT) scans of the chest.

Similar to mammography, a chest x-ray image provides only a 2D projection of the lung in which various structures overlap and lead to difficulties in interpretation. Even though the third dimension (3D) is added in CT imaging of the lung, difficulty in lung disease detection remains.

B. How lung CAD devices have been approved

1. Chest x-ray CAD

One chest x-ray CAD system was approved as a PMA due to its indications for use: it is trained and tested to identify malignancies.

The indications for the chest x-ray CAD state that this device identifies and marks regions of interest (ROIs) on digitized frontal chest radiographs. It identifies features associated with solitary pulmonary nodules from 9 to 30 mm in size, which could represent early-stage lung cancer. The device is intended for use as an aid only after the physician has performed an initial interpretation of the radiograph.

This PMA submission provided the following to support safety and effectiveness:

1. Chest radiographs were selected from a 1970s lung cancer screening trial database including male heavy smokers, 45 years of age or older, with a high risk for cancer. A set of 240 study cases, consisting of 80 cancer cases and 160 cancer-free cases, were selected from over 10,000 screening cases. The 80 cancer cases were primary lung cancers, from different patients, with lesions 9.5 to 27.5 mm in size, each proven pathologically by biopsy with location confirmed by a panel of radiologists.

2. Device standalone performance without readers was estimated.

3. A reader study, using receiver operating characteristic (ROC) analysis.

4. Fifteen (15) board-certified radiologists participated in the study.

5. Reader reviewed case three times over two separate sessions: a session without CAD assistance (independent read) and another session for the two other case reviews, i.e., so-called sequential read (first without CAD assistance and then immediately thereafter with CAD assistance).

6. Primary analysis used the area under the ROC curves and demonstrated that readers with the CAD increased their ROC performance for primary lung cancers up to 30 mm in diameter. Secondary analysis demonstrated that readers with the CAD increased their ROC performance for primary lung cancers up to 30 mm in diameter that had previously been missed at original screening interpretation, as well as their ROC performance for primary lung cancers 9 to 15 mm in average diameter (smaller cancers).

7. Analyses of the sensitivity and false positive fraction were provided.

8. ROC curves were all calculated without taking into account whether or not a cancer identified on the chest radiograph was attributed to the correct location. (As with any device that improves sensitivity for detecting some disease, there is a potential for harm to the patient from a concomitant increase of false positives. They could result in additional work-ups and possible interventions for lesions that turn out to be benign, or the false positive marks could be a distraction from true positives.)

2. CT lung CAD

Two CT lung CAD devices are currently being marketed. The indications for these devices state that these devices assist radiologists in the detection of solid pulmonary nodules during

review of multidetector CT (MDCT) scans of the chest. They are intended to be used as an adjunct to alerting the radiologist to regions of interest (ROIs) that may have been initially overlooked. They are intended to be used as a second reader after the radiologist has completed his/her initial read.

The applications for these CT lung CAD devices provided the following:

1. Chest CT scans were retrospectively collected from multiple clinical sites in various regions across the U.S.

2. Details on the study database:
   a. one submission had 151 cases: 63 nodule-present cases which were documented with a diagnosis of cancer, either primary lung cancer or an extrathoracic neoplasm (146 nodules identified), and 88 non-nodule cases;
   b. the other submission had 196 cases, all from patients referred for routine assessment of clinically or radiographically known or suspected pulmonary nodules: 181 cases had at least one nodule (1320 nodules identified), 15 cases had no nodule identified.

3. The clinical studies did not attempt to assess the effectiveness of the device on an asymptomatic lung cancer screening population.

4. Ground-truth was established by a consensus of a panel comprised of 3 (or up to 5) expert chest radiologists.

5. Up to 17 radiologists participated in the reader study.

6. Reader interpreted each case in a sequential read, first without CAD assistance (first interpretation) and then immediately thereafter with CAD assistance (second interpretation). Reader detections from the two interpretations were recorded separately. The reader task was to rate, both pre- and post-CAD, the probably of an actionable solid pulmonary nodule within each lung quadrant or lobe.

7. Primary analysis used the area under the ROC curves with the lobe or lung quadrant as a unit of analysis, and demonstrated that the average reader with CAD increased his/her area under the ROC curve for detecting lung quadrant/lobes containing a solid pulmonary nodule. Secondary analysis was performed on the patient level. Reader sensitivity and specificity were further analyzed.

8. Device standalone performance was derived and further analysis done per nodule size, per nodule shape, per nodule location in the lung, per patient protocol, etc.

FDA currently considers lung CAD devices indicated to assist the radiologist in identifying regions of interest that may have been overlooked as suitable for review as 510(k)s. However, devices with novel technologies or expanded indications for use may require PMA approval.

C. Published literature

1. Chest x-ray CAD

The published literature on chest x-ray CAD, is very limited and no data have been reported on the actual use of the device in the clinical practice environment,

The study by Berbaum et al.,[193] whose objective was to determine whether an idealized CAD, i.e., a CAD having no false positive detection and only cueing on an *added* simulated nodule, can reduce satisfaction-of-search effects in chest radiography (satisfaction-of-search is defined as occurring when a lesion is "missed" after detecting another lesion in the same image) effects in chest radiography, reported that:

- CAD prompts, even those that always point to their target lesion without false positive error, fail to counteract satisfaction-of-search in chest radiography;
- CAD prompts may induce less visual search for native abnormalities.

The study by Li et al.[194] was intended to determine retrospectively the sensitivity of and number of false-positive marks made by a commercially available computer-aided detection (CAD) system for identifying lung cancers previously missed on chest radiographs by radiologists, with histopathologic results as the reference standard. The term *missed cancer* was defined as including lesions that were mis-detected, and lesions that were detected but misinterpreted by radiologists. The CAD program was reported with an overall sensitivity of 35% for detection of missed lung cancers and an average of 5.9 false-positive marks per radiograph. The study concluded that the CAD system is able to detect a substantial proportion of visually subtle lung cancers that are likely to be missed by radiologists. The study may be of interest to the community as a baseline, but did not attempt to determine how the radiologists' performance might be influenced by the CAD results. The authors concluded that "the actual effects of CAD on radiologist accuracy and productivity in routine clinical practice remain[ed] to be determined, and further studies are required."[194]

Various academic papers reported standalone performance of CAD devices, including research CAD systems and commercial CAD devices, and proposed individual techniques to further improve the algorithm detection performance.[195,196] These papers used limited datasets of chest radiographs and shows varying CAD standalone performance ranging from 68.3% to 81.3% for the sensitivity, and from 4.5 to 1.4 average number of false positives per image.

2. CT lung CAD

The development of CAD devices for CT lung is relatively recent. The number of papers and studies in the peer reviewed literature is thus limited. The peer reviewed literature mostly reports CAD standalone performance improvements by academia and research groups, but no large reader studies. However, limited reader studies have been reported. They show a potential increase in reader sensitivity, especially for small lung nodules, and a potential use in relation to assessing nodule size (or other characteristics) on serial CT scans over time.[197,198,199,200]

D. Clinical testing issues specific to chest x-ray and CT lung CAD

1. Ground truth

Four potential techniques have been identified for determination of the ground truth for lung CAD devices. They consist of:

- expert panels
- expert panels and follow-up CT scan of the patient (may allow verification of normal patients and an identification of higher suspicion nodules if they increase in size over time)
- expert panels and PET/CT
- pathology

Biopsy-proved pathological diagnosis would be an ideal standard for ground truth definition and precise diagnosis of the nodule (and thus patient), if available. However, pathology might not be completely sufficient, even for nodule cases, because the location of the nodule in the chest x-ray or CT scan would still need to be determined by an expert panel. More importantly, pathology is not available for many nodules identified on chest x-ray or CT.

An alternative is to have a panel of radiologists (different from the radiologist participating in the evaluation of the device) identify all appropriate pulmonary nodules on the chest x-ray or CT scans. The number of expert should be large enough to enable adequate sampling of reader variability for this task. One suggest approach has been developed to identify, as completely is possible, all nodules in a scan without requiring forced consensus.[201] Panel review is conducted in two phases.[201] An initial "blinded read phase" allows the radiologists to independently mark lesions meeting a pre-established criteria. A subsequent "unblinded read phase" is then conducted. In this second phase results from all radiologists are revealed to each of the radiologists, who then independently review their marks along with the marks of their colleagues. A radiologist's own marks could be left unchanged, deleted, or switched in terms of lesion category, or additional marks could be added. A possible alternative would be to identity lesions with the use of a combined PET/CT exam if this data is available. Note that patients undergoing this type of exam may not represent the general population of patients having Thoracic CT examinations.

If precise diagnosis of each nodule is necessary for CAD devices intended to identify specific pathologies (e.g., cancerous nodules), pathology on all nodules and follow-up for normal patients should be available.

Topic for discussion

L1. In light of FDA's least burdensome requirements, please discuss the merits and practicality of possible options for determining ground truth for lung CAD devices.

2. Standalone performance testing

Standalone performance may include sensitivity, the number of false positives per scan and/or the FROC performance curve for the device. Standalone performance is based on a dataset of chest x-ray or CT cases that are distinct from the dataset used to train the device. Both sensitivity and the FP rate may be important measures of performance as they provide information to the radiologist on the strengths and weaknesses of these devices in identifying nodules and data for labeling the device, as well as providing an estimated performance measure as baseline comparison for any subsequent modification to the CAD algorithm. The FROC curve is a plot of the sensitivity vs. the number of FP marks for a full range of possible thresholds applied to the CAD output score.

The detection performance of lung CAD devices depends on a number for factors including nodule size, location, shape (spherical, irregular), pathology, opacity (solid, semi-solid); CT system, slice thickness, patient dose, and imaging protocol; and co-morbidities affecting lung imaging. Stratification of standalone performance data may help to assess the performance and potential utility of lung CAD devices. In this case the testing database for device standalone performance may require a database representative of various substrata in order to provide meaningful information on device capabilities and correct labeling.

Topic for discussion

L2. What substrata, if any, should be considered in standalone performance testing and labeling of lung CAD devices?

3. Reader performance testing

Standalone performance testing indicates how well the device marks locations of interest. However, standalone performance does not indicate what locations will be marked by a physician using the CAD device. As discussed in Section II.C.6., physician performance can be measured using a prospective study, or a retrospective study, such as an MRMC study design, which may be considered ‘efficient’ in terms of the number of readers and cases. This methodology accommodates statistical analysis for a number of common performance metrics (e.g., ROC, AUC, sensitivity, specificity). Other clinical study designs could also be considered.

Topics for discussion

L3. Are both device standalone performance and reader performance testing necessary to demonstrate that lung CAD devices provide clinically significant results? Would one or the other suffice? Are there other types of studies that should be provided instead or additionally?

L4. Chest x-ray and chest CT are done for many important reasons other than looking for lung nodules. Can the use of CAD affect the diagnosis for these other conditions? Can the presence of other conditions alter the effectiveness of the CAD function or the risk-benefit profile of a lung CAD device?

Test dataset

Chest x-ray is usually used to diagnosis the potential causes of symptoms such as shortness of breath, a bad or persistent cough, chest pain, injury, or fever. The vast majority of chest x-ray exams indicated for patients with symptoms unrelated to lung nodules will likely not contain lung nodules. For this reason, an enriched database might be used. For CADs intended for any chest x-ray examination, an enriched database might include:

- a majority of cases with no nodules
- most nodule cases containing only a single nodule with a range of nodule sizes.

For the chest x-ray CAD indicated specifically for suspicion of diseases likely to produce lung nodules, (e.g., suspicion for cancer or metastatic cancers), the testing database might include a:

- minimal number of cases with no nodules
- similar number of cases containing single and multiple nodules
    - cases with a single nodule have a range of nodule sizes
    - cases with pulmonary metastases (multiple nodules).

Chest CT is usually performed on patients in whom an abnormality is suspected (e.g, from a previous chest x-ray examination). It could be for abnormalities other than those of the lung. The review/interpretation of these CT scans will still include a check of the lung field. If the abnormality is located in the lung, CT is to: help diagnose clinical signs or symptoms of disease of the chest, detect and evaluate the extent of tumors that arise in the lung and mediastinum, or tumors that have spread there from other parts of the body, or to assess whether tumors (not necessary in the lung) are responding to treatment, or to help plan radiotherapy (not necessary in the lung). Again, en enriched database might be appropriate. The testing database for a CT lung CAD intended for use on all patients referred to chest CT might include:

- a majority of cases with no nodules
- most cases with nodules containing only a single nodule with a range of nodule sizes

However, for chest CT CADs indicated specifically for suspicion of disease likely to produce lung nodules (e.g., suspicion for cancer or metastatic cancers), the testing database might include:

- minimum number of cases with no nodules
- similar number of cases containing single and multiple nodules
    - cases with a single nodule have a range of nodule sizes but the majority small
    - cases with pulmonary metastases (multiple nodules).

If a CT lung CAD is intended for use on patients already diagnosed with lung cancer (intended for use during therapy), the testing database might include:

- Similar number of cases containing single and multiple nodules
    - cases with a single nodule having a range of nodule sizes but the majority small
    - cases with extensive lung diseases (multiple nodules).

Topic for discussion

L5. If a lung x-ray or CT CAD is intended for use on all patients referred to chest x-ray or CT examination, is an enriched, retrospective reader study or a stress test study an acceptable alternative to a field test? If so, are the following the appropriate characteristics to be considered when designing an enriched database or stress test? Are there other characteristics that should be considered? Please discuss alternatives for this and other target populations.
- number of patients with no nodules, single nodules, or multiple nodules
- range of nodule sizes

Study endpoints

Various endpoints can be utilized in the clinical assessment of lung CADs. To use ROC-based endpoints, the study would require using a rating or sorting that is not typically used in clinically practice. This rating may be a subjective confidence regarding the presence or absence of a nodule or a similar type of rating. The scale used should be well defined, and include multiple categories (or a continuous scale). Reader training would be necessary to familiarize the physicians with both the scoring scale and reading protocol. If the ROC paradigm is used, location accuracy may not be accounted for in the analysis. An alternative could be to use a region-based (e.g., quadrant or lobe) assessment instead of the by-patient assessment. This has the advantage of increasing the statistical power of a given set of patients, but would not provide a direct measure of the effect of the device use on individual patients. Note, a direct assessment of the individual patient could be provided alone with the regional rating or an indirect measure of patient-based performance could be produced with a combination of the region ratings. With either protocol, it may be important to establish that the precision in reader localization is adequate. This could include some type of by lesion or another type of assessment in addition to any by patient or by region analysis. This additional assessment would provide confidence that the readers are accurately localizing true lesions in the study.

Topics for discussion

L6. Please discuss the merits and practicality of ROC analysis (full curve, area under the curve, partial area under the curve, sensitivity and specificity) as a primary endpoint(s). Are there alternatives? If so, please discuss their merits. What are clinically significant effect sizes?

L7. Please discuss the merits of per patient, per region, and per lesion endpoints for testing of lung CAD devices. If per patient analysis is used, should the study endpoints also include an assessment of location accuracy?

L8. CAD use can impact reading time. Should reading time be assessed? If so, how?

Reader paradigms

Kobayashi et al. showed that reader detection accuracy increases for both sequential and concurrent reader paradigm when using a particular chest x-ray CAD.[202] This suggests that chest x-ray CAD and potentially lung CT CAD may fit in the radiology workflow if implemented either as a sequential or concurrent reader. The main advantage of concurrent CAD over sequential CAD is that it may lead to a reduction in the reading time for interpretation of the CTC scan. This may also be true for sequential CAD, if the initial unaided read by the physician is not as exhaustive an interpretation of the scan as that used in a conventional interpretation without CAD.

Topic for discussion

L9. FDA does not specify indications for use, but reviews indications for use that are requested by Sponsors. What are the Panel's views regarding second reader versus concurrent reading using a CAD device? Specifically,

a. How are lung CADs for chest x-ray and chest CT used clinically?
b. Are second reader and concurrent reading modes both clinically relevant options for use in practice? If not, which paradigm(s) are appropriate for lung CAD devices?
c. Do you believe users understand that if a device is labeled as a second reader, the (i.e. the physician) should always read the radiological image completely before turning on the CAD?

V. Colon CAD devices

A. Introduction of Colon CAD into the clinical practice

1. Background

Colorectal cancer is a leading cause of cancer mortality in U.S. with over 52,000 deaths estimated to have occurred in 2007.[203] It is now well-established that screening for colon cancer will reduce both mortality and incidence.[204, 205] Most colorectal cancers are believed to arise within adenomatous polyps.[206, 207] Nonadenomatous polyps (of which hyperplastic polyps account for the majority) are not considered to have malignant potential.[208] It is not currently possible to distinguish adenomatous from non-adenomatous polyps on visual inspection. The probability of a polyp being adenomatous is strongly dependent on size[209] as most subcentimeter polyps are not adenomatous. This suggests that polyp size is an important factor to be used in determining the need for colonoscopy and/or polypectomy. The advanced adenoma (defined as an adenoma with: a size of at least 10 mm; or the presence of a substantial villous component; or the presence of high-grade dysplasia[210]) represents the optimal target lesion for removal at colonoscopy.

Guidelines recommend the screening of adults who are at average risk for colorectal cancer (i.e., all persons above the age of 50).[211] Recommendations as to the frequency of screening remain variable and are also determined by findings at prior colonoscopy. A large fraction of the average-risk population in the United States does not follow recommended screening procedure for this largely preventable disease, because of perceived inconveniences and the potential discomfort of existing screening tests.[212]

2. Standard of care

Current screening procedures for colonic polyps and colorectal cancer include:

- Barium enemas

  A fluoroscopic examination whereby a rectal tube is inserted and barium fills the colon under gravity (with a double contrast examination air is also insufflated) and 2D x-ray images are obtained. Polyps appear as filling defect on single contrast examination, or are shown in relief on double contrast examination. Barium enema is no longer commonly used for screening in clinical practice.

- Fecal occult blood testing

  A laboratory test whereby blood contained in a stool sample is identified by chemical reactions with specific reagents. This test is commonly used to identify many clinical conditions associated with blood in stool (not specific to colon cancer) because it is inexpensive and easy to perform in a doctor's office.

- Optical colonoscopy (OC)

  An examination whereby a fiberoptic scope is inserted through the rectum and provides direct visualization of the entire colon. Carbon dioxide (or air) is insufflated to distend the colon during the examination. Patients undergo bowel cleansing prior to the examination to eliminate residual stool that can interfere with the visualization of the endoluminal surface of the colon. Patients are typically given conscious sedation.

  Optical colonoscopy is well-established as screening examination for colonic polyps and colorectal cancer in the U.S. and used both for identification of polyps (and thus detection of advanced colorectal neoplasia) and for local resection. Approximately 1.5 to 2 million screening OCs are performed each year in the U.S.

3. CT Colonoscopy

With advances made in imaging technologies, computed tomography (CT) offers the possibility to provide, with great clarity for medical assessment, images of internal organs, bone, soft tissue and blood vessels. CT applications are covering all body parts and thus include acquisition of the lower thorax, abdomen, and pelvis. Vining et al.[213] introduced computed tomography colonoscopy (CTC, or CT Colonography) in 1994 as an alternative to barium enemas. It was further indicated as a first-line alternative to screening colonoscopy.[214] CTC is often referred as virtual colonoscopy (VC) because data from computed tomography (CT) are acquired and processed to generate 2D and/or 3D displays of the colon and rectum.[215, 216] Virtual colonoscopy has shown promise as a method of colorectal cancer screening that may be acceptable to many patients.[217]

Computed tomography (CT) systems are not cleared for any screening indications, but current medical practice uses CTC to identify patients with colonic polyps (size, location and number) and can be used to help guide optical colonoscopy. Kim et al.[215] conducted a comparison study involving 3120 consecutive adults in a primary CTC screening interpretation arm and 3163 consecutive adults in a primary OC screening interpretation arm, and showed that CTC and OC screening strategies resulted in similar detection rates for advanced neoplasia. Pickhardt et al.[214] conducted a similar comparison study involving 1233 asymptomatic patients. The study also showed that CTC with the use of a 3D primary review is an accurate screening method for the detection of colorectal neoplasia in asymptomatic average risk adults and compares favorably with optical colonoscopy in terms of the detection of clinically relevant lesions.[214] Other studies have found CTC to be less sensitive than OC for detecting clinically significant polyps using 2D reading paradigms.[218] Reading mode, patient preparation and CT protocol all contribute to this reported differences in CTC performance.

On October of 2007, ACRIN released trial results for ACRIN Protocol 6664 (available at: http://www.acrin.org/6664_protocol.html), The National CT Colonography Trial. This trial used a prospective design involving 2,531 patients at 15 U.S. clinical sites. The trial did not involve the use of CAD. However, it provided estimates of reader sensitivity and specificity in polyp detection and the data can therefore be used to identify polyp and/or

patient characteristics that may colonic polyps more difficult for radiologists to detect at CT. Table 4 is a copy of these estimates per each polyp subgroup. Specificity remains high, 86% to 89%, across all lesion sizes. The sensitivity was dependent on the lesion size.

Table 4: Radiologist accuracy for detecting patients with polyps ranging in size from 5 mm or larger to 10 mm or larger from ACRIN Protocol 6664

	≥ 5 mm	≥ 6 mm	≥ 7 mm	≥ 8 mm	≥ 9 mm	≥ 10 mm
Sensitivity	65%	78%	84%	87%	90%	90%
Specificity	89%	88%	87%	87%	86%	86%
PPV	45%	40%	35%	31%	25%	23%
NPV	95%	98%	99%	99%	99%	99%

4. Requirements for CTC interpretation

Patient preparation, CT protocol and procedure

Patient preparation is an important concern for accurate evaluation of the colon. Similar to OC, CTC examination requires bowel cleansing. Patients often consume oral contrast agents (barium-containing and/or water soluble) to mark solid stool and for the opacification on luminal fluid.[219] Prior to the CT scan, an enema tip or catheter is placed in the rectum and the colon is filled with air or carbon dioxide until the patient reports a sensation of fullness. Few patients report significant discomfort. Both supine (patient lying on their back) and prone (patient lying on their stomach) CT images are acquired to help distinguish polyps from any residual stool. Current multi- detector CT scanners can acquire all the image data in two breath holds (one each for supine and prone), minimizing motion artifacts. The entire procedure is complete within 15-20 minutes. Patient sedation is not required but can be utilized.

Radiologist interpretation

The radiologist examines 2D axial images of the entire length of the colon as well as 3D reconstructions to view any part of the colonic lumen from any angle that the radiologist selects. An advanced viewing workstation also provides endoluminal fly-through capability (i.e., virtual colonoscopy). It is however important to keep in mind that CTC examinations must also be carefully reviewed for extracolonic findings. It is also important for the interpreting radiologist to remain cognizant of the diagnostic limitations imposed by the absence of intravenous contrast material and the reduced dose of radiation that is sometimes employed. Interpretation time for the average CTC studies is about 10-15 minutes.

CTC reporting

Similar to the ACR BI-RADS for reporting mammography, the C-RADS reporting system has been proposed (although not universally used) by Zalis et al.[220] The C-RADS

utilizes both polyp size, number of polyps, CT attenuation, and lesion morphology to determine the assessment category for patient management as illustrated below.

Lesion Size (mm)—For lesions 6 mm or larger, the single largest dimension of the polyp head (excluding stalk if present) on either multiplanar reconstruction (MPR) or 3D views. The type of view employed for measurement should be stated.

Morphology
- Sessile—broad-based lesion whose width is greater than its vertical height
- Pedunculated—polyp with separate stalk
- Flat—polyp with vertical height less than 3 mm above surrounding normal colonic mucosa

Location
- Refer to named standardized colonic segmental divisions: rectum, sigmoid colon, descending colon, transverse colon, ascending colon, and cecum.

Attenuation
- Soft-tissue attenuation
- Fat

Polyp size is the critical measure that determines the probability that a polyp will develop into a malignancy as well as the need of immediate OC or follow-up imaging/OC. Size evaluation is especially crucial for screening with CTC because smaller lesions may be identified but left in situ for future surveillance. The size of the polyp can be measured on magnified two-dimensional multiplanar images. The size of irregularly shaped lesions may be underestimated or overestimated on 2D/3D views. In cases of oval polyps, perhaps measured more easily on three-dimensional endoluminal views, but three-dimensional measurement of size is subject to potential distortion. For standardization, the largest single dimension should be reported for a given lesion. In investigational studies, it is important to specify how measurements are performed and set rules for uniformity between the radiologists and the endoscopists.[221]

The majority of polyps 5 mm or smaller are hyperplastic and are not thought to confer increased risk for development of colon carcinoma. Detection of even one polyp greater than or equal to 10 mm results in a referral to colonoscopy for polypectomy. Studies report 90% sensitivity for detection of such polyps by radiologists. Reader sensitivity is decreased for detection of smaller polyps. The number of polyps also helps determine patient management. When three or more synchronous 6–9-mm polyps are detected at CTC, referral to colonoscopy for polypectomy is currently recommended.

Reporting the final diagnostic assessment for the patient supplements assessment on each finding. The diagnostic assessment for the patient is separated into 5 categories, as follows:

C0. Inadequate Study/Awaiting Prior Comparisons
- inadequate prep: cannot exclude lesions ≥ 10 mm owing to presence of fluid/feces
- inadequate insufflation: one or more colonic segments collapsed on both views
- awaiting prior colon studies for comparison

C1. Normal Colon or Benign Lesion; Continue Routine Screening*1
- no visible abnormalities of the colon
- no polyp ≥ 6 mm
- lipoma or inverted diverticulum
- nonneoplastic findings—eg. colonic diverticula

C2. Intermediate Polyp or Indeterminate Finding; Surveillance or Colonoscopy Recommended*2
- intermediate polyp 6-9 mm, < 3 in number
- indeterminate findings, cannot exclude polyp ≥ 6 mm in technically adequate exam

C3. Polyp, Possibly Advanced Adenoma; Follow-up Colonoscopy Recommended
- polyp ≥ 10 mm
- ≥3 polyps, each 6-9 mm

C4. Colonic Mass, Likely Malignant; Surgical Consultation Recommended*3
- lesion compromises bowel lumen, demonstrates extracolonic invasion

*1: Every 5-10 years.

*2: Evidence suggests surveillance can be delayed at least 3 years, subject to individual patient circumstance.

*3: Communicate to referring physician as per accepted guidelines for communication, such as ACR Practice Guideline for Communication: Diagnostic Radiology. Subject to local practice, endoscopic biopsy may be indicated.

5. Origin of colon CAD devices

The high radiologist sensitivity for detection of polyps greater than or equal to 10 mm (90%) may leave little room for improvement. On the other hand, there is more room for improvement of radiologist sensitivity for detection of smaller polyps. This is a potential clinical niche for using CTC colon CAD. Another potential area of improvement is reduction in CTC interpretation time.

Topic for discussion

C1. Please discuss the potential clinical utility of CTC Colon CAD. Things to consider include:
- improved sensitivity to detect polyps of different sizes
- reduced reading times
- guiding optical colonoscopy

B. Cleared colon CAD devices

At least two CTC colon software devices with early CAD-like functions have been cleared by FDA. The IFU statements were:

- to "confirm the presence/absence of physician identified polyps"
- "automated highlight feature for the visual identification of spherical structures within the colon"
- "produces images that highlight spherical anatomical regions, such as polyps, and/or stool"

CTC colon CAD devices that are only intended to assist radiologist in the detection of potential polyps (without claiming diagnosis, e.g., cancer), and are intended to be used for patients who would otherwise be referred to OC, may be Class II requiring a premarket notification (510(k)) submission. This classification is based on the intended use, the indications for use, and the risk related to use of these devices.

C. Published literature

Over the last five years, academia and the National Institutes of Health (NIH) have devoted research efforts to the development of colon CAD devices intended to detect polyps on CTC scans. Numerous studies have been published reporting preliminary device standalone performance as well as improvement over time, and retrospective reader studies to evaluate the impact of the device use under various reading paradigms. Studies also focused on the effect of the reading methodology, i.e., 2D versus 3D. Below is a summary of some of these studies.

CT colon CAD standalone performance

Preliminary studies of prototype CAD systems on small patient datasets have reported per-polyp sensitivities from 64% to 100% and false-positive rates from 1 to 11 false positives per patient for detecting polyps 1 cm.[212, 222, 223, 224, 225, 226, 227]

Summer et al.[212] reported standalone performance results of their investigational CTC-CAD using a large, consecutive, prospectively enrolled asymptomatic screening patient population. The test dataset consisted of 792 asymptomatic patients undergoing CTC. The per-polyp and per-patient sensitivities for CAD were both 89.3% (25/28; 95% confidence interval, 71.8%–97.7%) for detecting retrospectively identifiable adenomatous polyps at least 1 cm in size. The false-positive rate was 2.1 (95% confidence interval, 2.0 –2.2) false polyps per patient. Summers et al. concluded that their per-patient sensitivity of CT virtual colonoscopy CAD in an asymptomatic screening population is comparable to that of optical colonoscopy for adenomas >8 mm.

Yoshida et al.[222] reported assessment of their investigational CAD device. The study included 71 patients: 14 had colonoscopically confirmed polyps (a total of 21 polyps). Of the 21 polyps, 15 were at least 5 mm but less than 10 mm in diameter. The remaining six polyps were 10 mm or more in diameter, one polyp measured 25 mm, and the other five polyps were less than 20 mm. The size of the polyps was determined primarily on the basis of the colonoscopy and pathology reports. In the per patient analysis, sensitivity was 100%, with an average false-positive rate of 2.0 per patient. In the per polyp analysis, the device detected 90% of the polyps at the same false-positive rate.

A study by Taylor et al.[228] was to assess the sensitivity of computer-assisted reader software for polyp detection and further compare the performance with that of expert reviewers. The 25 test dataset consisted of 25 cases and included a total of 32 polyps ranging from 6 to 35 mm in diameter. The software identified 81% of all polyps, 92% of polyps ≥ 10 mm which compared with an average sensitivity of 70% for the expert reviewers. All polyps missed by experts 1 (n = 4) and 2 (n = 3) and 12 (86%) of 14 polyps missed by expert 3 were detected by the device. The median number of false-positive highlights per case was 13, which 91% were estimated as easily dismissed.

Retrospective reader studies

Shi et al.[229] investigated a retrospective evaluation of the effect of three-dimensional (3D) viewing on radiologists' accuracy in classifying true positive (TP) and false-positive (FP) polyp candidates identified with computer-aided detection (CAD). Forty-seven CTC data sets were used in the study. Four radiologists classified 705 polyp candidates (53 TP candidates, 652 FP candidates) identified with CAD; initially, only 2D images were used, but these were later supplemented with 3D rendering. Another radiologist unblinded to colonoscopy findings had to characterize the features of each candidate, assessed colon distention and preparation, and defined the true nature of FP candidates. The study showed that the use of 3D viewing increases reader accuracy in the classification of polyp candidates identified with CAD. The accuracies depended significantly on the polyp size and examination quality.

Taylor et al.[230] compared primary 3D endoluminal analysis with primary 2D transverse analysis supplemented by CAD software used concurrently for CTC polyp detection and reader reporting times. The dataset cases were restrospectively selected and consisted of 20 CTC scans from 14 patients. There were a total of 48 polyps endoscopically proved. Polyp coordinates were documented in consensus by three unblinded radiologists. Two radiologists read the data sets, which were randomized between primary 3D endoluminal views with 2D problem solving and 2D views supplemented by CAD device. Mean sensitivity values for polyps measuring 1–5, 6–9, and 10 mm or larger were 14%, 53%, and 83%, respectively, for 2D CAD software analysis and 16%, 53%, and 67%, respectively, for primary 3D analysis. Overall sensitivity values were 41% for 2D CAD software analysis and 39% for primary 3D analysis. One of the readers detected significantly more polyps than the other reader, particularly when using the 3D flythrough method. Mean false-positive findings were 1.5 for 2D analysis and 5.5 for 3D analysis. The reading times were found significantly increased with the 3D method.

Petrick et al. [231] investigated the effect of a CT colon CAD as second reader on radiologists' diagnostic performance in interpreting CTC examinations for a primary 2D reading approach with 3D problem solving. They had four board-certified radiologists read first without then with CAD in a sequential reader study design. The study consisted of 60 CT examinations. Interpretations of these cases were performed using a commercially available review system. The results showed non-significant reader AUC increases of 0.03, 0.03, and 0.04 for patients with adenomatous polyps 6 mm or larger, 6–9 mm, and 10 mm or larger, respectively. However, CAD was found to increased reader sensitivity by 0.15, 0.16, and 0.14 for those respective groups, with a corresponding decrease in specificity of 0.14.

Halligan et al.[232] performed a similar reader study. Ten readers trained in CT but without special expertise in colonography interpreted CTC images of 107 patients (60 with 142

polyps), first without CAD and then with CAD after temporal separation of 2 months. Per-patient and per-polyp detection were determined by comparing responses with known patient status. With CAD, 68% (95% confidence interval [CI]: 55%– 80%) of patients with polyps were identified more frequently by readers. Per-patient sensitivity increased significantly in 70% of readers, while specificity dropped significantly in only one. Polyp detection increased significantly with CAD; on average, 9.1% (95% CI, 5.2%–12.8%) of polyps were detected by each reader. Small- (≤5 mm) and medium-sized (6 –9 mm) polyps were significantly more likely to be detected when prompted correctly by CAD. However, overall performance was relatively poor; even with CAD, average readers detected only 51.0% of polyps with size ≥10 mm, and 38.2% of polyps with size ≥6 mm. Interpretation time was shortened significantly with CAD: by 1.9 minutes (95% CI, 1.4 –2.4 minutes) for patients with polyps and by 2.9 minutes (95% CI, 2.5–3.3 minutes) for patients without. The authors concluded that overall, 90% of the readers benefited from CAD, either by increased sensitivity and/or by reduced interpretation time.

D. Clinical testing issues specific to colon CAD

1. Ground truth

Several techniques have been identified as possible methods for ground truth determination (polyp presence, location and extent, matching between the supine and prone scans) for CTC datasets used for standalone performance and clinical performance testing. They consist of:

- optical colonoscopy (OC)
- optical colonoscopy (OC) and CTC with segmental unblinding
- optical colonoscopy (OC) and a panel of experts reviewing the CTC (without the CAD device to be tested) to include polyps missed at OC

Determination of the presence, location and extent of polyps:

Optical colonoscopy (OC) and CTC with segmental unblinding as used in CTC efficacy trials may be the optimal method for ground truth determination. In such trials, both normal patients and diseased patients (i.e., patients with polyps and recommended to undergo OC examination for resection) undergo OC examination. The CTC scans of each patient are interpreted prior to OC. In the OC procedure, each segment of the colon is first evaluated without any knowledge of the CTC results. The CTC results are then revealed to the gastroenterologist who reevaluates the segment when necessary to confirm the presents of any CTC identified lesions. This process is repeated for each segment of the colon. This provides a high level of truth determination reducing the number of false negative finding because of the independent application of CTC and OC. It also provides a higher level of truth on normal patients, again because patient status is confirmed by two different evaluations. This ground truth determination may be difficult and costly to implement requiring the consent from the patient prior to his/her participation to such testing.

An alternative technique for determination of ground truth would be to use OC and a panel of experts reviewing the CTC (without the CAD device to be tested) to include polyps missed at OC as well as to confirm normal patients and to identify polyps on patients that were determined to have no more than two synchronous 6-9 mm polyps. When CTC reveals at least one polyp greater than or equal to 10 mm, or three or more synchronous 6–9-mm polyps, patients are recommended to undergo OC examination for resection of these polyps as well as other polyps that may be revealed during the procedure. Therefore, polyps detected at CTC can be confirmed during OC examination as being either true positives or false positives. This makes OC a natural choice for ground truth determination in this case. OC is known to have a high sensitivity for detecting polyps, but as shown in the Pickhardt et al.[214] study, OC is not perfect and misses a small fraction of these polyps. The use of only OC would produce a fairly strong certainty in the ground truth for polyps, especially for advanced adenomas, with the knowledge that some fraction of false negatives would have been missed. In case of CTC interpretation reported as either normal or as containing no more than two synchronous 6-9 mm polyps, patient would not be recommended for OC. Therefore, ground truth determination for these cases may not be provided by OC. An alternative for ground truth determination would be to have a panel of expert radiologists confirm normal patients and identify polyps on patient that were identified with no more than two synchronous 6-9 mm polyps. The addition of a follow-up CTC examination, if available, could further allow confirmation of the ground truth determination by the panel of experts, thus lowering uncertainty introduced by the panel of experts.

Determination of polyp matching between supine and prone scans:

While OC is a good tool to confirm that polyps exist, it is not always straightforward to make a one-to-one correspondence between a polyp visible on OC and that same polyp on the CTC exam. Likewise, matching polyps between supine and prone exams can pose significant difficulties. It is possible that a panel of physicians could be used in any of the truthing methods discussed above to both match OC and CTC polyps and match polyps between the supine and prone exams. Multiple readers would likely be required in the ground truth determination so that inter-reader variability in interpretation could be accounted for in the overall assessment.

Topic for discussion

C2. In light of FDA's least burdensome requirements, please discuss the merits and practicality of possible options for determining ground truth for lung CAD devices.

2. Standalone performance testing

Standalone performance may include sensitivity, the number of false positives per scan and/or the FROC performance curve for the device. Standalone performance utilizes a dataset of CTC cases that are distinct from the dataset used to train the device. Both sensitivity and the FP rate are important measure of performance as they provide information to the radiologist on the strengths and weaknesses of these devices in polyp detection, data for labeling the device as well as providing an estimated performance measure as baseline comparison for any subsequent modification to the CAD algorithm.

The FROC curve is a plot of the sensitivity vs. the number of FP marks for a full range of possible thresholds applied to the CAD output score.

The detection performance of colon CAD devices may depend on a number for factors including the polyp size, pathology (e.g., adenoma, hyperplastic), type (sessile, flat, or pedunculated); CT system, patient dose, and imaging protocol; and patient preparation and the use of fecal tagging. Stratification of standalone performance data may provide useful information regarding the performance and potential utility of these devices in these substrata.

Topic for discussion

C3. What substrata, if any, should be considered in standalone performance testing and labeling of lung CAD devices?

3. Reader performance testing

Standalone performance testing indicates how well a CTC colon CAD device marks locations of interest. However, standalone performance does not indicate what locations will be marked by a physician using the CAD device. As discussed in Section II.C.6., physician performance can be measured using a prospective study, or a retrospective study, such as an MRMC study design, which may be considered 'efficient' in terms of the number of readers and cases. This methodology accommodates statistical analysis for a number of common performance metrics, e.g., ROC, FROC, sensitivity, specificity. Other clinical study designs could also be considered.

Topic for discussion

C4. Are both device standalone performance and reader performance testing necessary to demonstrate that colon CAD devices provide clinically significant results? Would one or the other suffice? Are there other types of studies that should be provided instead or additionally?

Test dataset

Possible options for a test dataset for a clinical study are to:

1) retrospectively collect cases using a prospectively developed protocol so that the case mix represents the overall screening population

2) select cases from a database with a significantly greater percentage of abnormal patients producing an enriched set representative of the true population

3) select cases from a database of difficult cases that would be expected to challenge the readers, producing a stress test set of cases used to establish where the CAD is most beneficial.

An example of a test dataset for a stress test study of a CT colon CAD device might include a population of asymptomatic patients 50 years or older with enrichment such that:

- approximately 50% of patients have polyps and 50% are normal
- patients with polyps have predominantly 1 or 2 polyps
- polyps are predominantly in the 6-10 mm size range
- powered to show statistical significance for the defined endpoint.

As an example, if the statistically significant endpoint required 200 patients, then these criteria might be met from a population indicated below (**Figure 10**). The actual sample size would need to be adjusted according to the hypothesis to be tested, the statistical plan, and the number of readers.

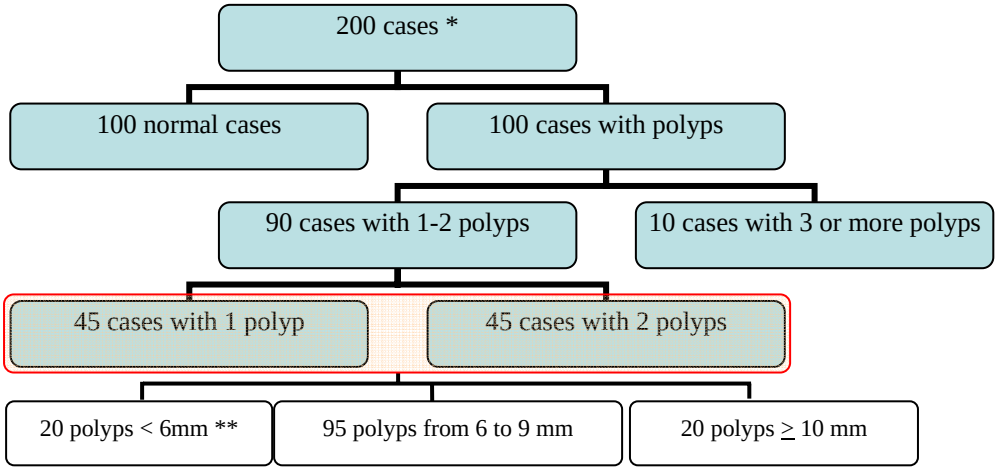


Figure 10: Example of the Database Partitioning for test dataset for the evaluation of of colon CAD devices. *The actual sample size would need to be adjusted to hypothesis, statistical plan, and appropriate number of readers; a pilot study is recommended for correct estimating this size. ** Appropriate size cut-offs may depend on the variability in measurement of polyp sizes.

Topic for discussion

C5. Is an enriched, retrospective reader study or a stress test study an acceptable alternative to a field test? If so, are the following the appropriate characteristics to be considered when designing an enriched database or stress test? Are there other characteristics that should be considered?

- proportion of patients having polyps
- proportion of patients with multiple polyps
- polyp size

Study endpoints

CTC is being used to identify patients who should undergo short-term follow-up imaging, OC examination, or surgery. This suggests that a per patient assessment may be appropriate in a clinical study.

If an ROC analysis is used, a rating or ordering would be needed; e.g. the rating might be a subjective confidence regarding the presence or absence of a polyp. The scale used should be well defined, and include multiple categories (or a continuous scale). Such a rating is not typically used in clinical practice; hence reader training is generally needed to familiarize the physicians with both the scoring scale and reading protocol.

If the ROC paradigm is used, location accuracy might not be accounted for in the analysis. An alternative could be to use a region or segment-based assessment instead of per patient assessment. This has the advantage of increasing the statistical power of a given set of patients, but would not provide a direct measure of the effect of the device use on individual patients. However, a direct assessment of the individual patient could be provided along with the regional rating, or an indirect measure of patient-based performance could be produced with a combination of the region ratings.

With either protocol, per lesion analysis can be used to establish that the precision in reader localization is adequate. This per lesion analysis could be in addition to any per patient or per region analysis. This additional assessment could provide confidence that the readers are accurately localizing true lesions in the study.

Questions to the panel

C6. Please discuss the merits and practicality of ROC analysis (full curve, area under the curve, partial area under the curve, sensitivity and specificity) as a primary endpoint(s). Are there alternatives? If so, please discuss their merits. What are clinically significant effect sizes?

C7. Please discuss the merits of per patient, per segment, and per lesion endpoints for the colon CAD devices. If per patient analysis is used, should the study endpoints also include an assessment of location accuracy?

C8. CAD use can impact reading time. Should reading time be assessed? If so, how?

Reader paradigms

The published literature suggests that CT colon CAD could fit in the radiology workflow if implemented either as a sequential or concurrent reader. The main advantage of concurrent CAD over sequential CAD is that it may lead to a reduction in the reading time for interpretation of the CTC scan. This may also be true for sequential CAD, if the initial unaided read by the physician is not as exhaustive an interpretation of the CTC scan as that used in a conventional interpretation without CAD, or if the additional time necessary to review the CAD is small.

Topic for discussion

C9. FDA does not specify indications for use, but reviews indications for use that are requested by Sponsors. What are the Panel's views regarding second reader versus concurrent reading using a CAD device? Specifically,

a. How would colon CADs be used clinically?
b. Are second reader and concurrent reading modes both clinically relevant options for use in practice? If not, which paradigm(s) are appropriate for colon CAD devices?
c. Do you believe users understand that if a device is labeled as a second reader, the (i.e. the physician) should always read the radiological image completely before turning on the CAD?

In addition to the reading paradigm, the clinical impact of a colon CAD device could be tied to a number of patient, imaging and review factors specific to CTC imaging. These include many of the characteristics outlined in the standalone performance section above. Characteristics such as patient prep, stool tagging, 2D vs 3D fly-through primary review, CT parameters, etc. could impact the clinical aid provided by a colon CAD device.

Topic for discussion

C10. Are there other patient, imaging and review factors specific to colon CAD that should be considered in the assessment or labeling of these devices?

VI. Other CAD Devices

We have discussed issues relating to the demonstration of safety and effectiveness for three families of CAD devices—mammography CAD, lung CAD, and colon CAD devices. We will ask the panel to discuss the types of testing that are needed for each device type, e.g. standalone and/or reader studies, test datasets, study endpoints, ground truth, and reader paradigms.

In this section, we provide background information needed for the panel to consider how their recommended framework can be extended to other CAD devices that may be under development now or in the future.

A. Other CAD applications

We would like the panel's comments on the applicability of this meeting's discussions to to other types of CAD devices that may be developed in the future.

CAD could also be used:

- to search for cancer in other body parts
  - liver
  - pancreas
- with other imaging modalities
  - MRI
  - ultrasound
- for biopsy guidance
  - prostate
- to identify non-cancer abnormalities
  - acute ischemic lesions
- for monitoring responses to therapy or disease progression
  - response to cancer treatment
- to provide diagnostic assessment
  - diagnosing Alzheimer's disease from brain perfusion
  - early differential diagnosis

Topic for discussion

What emerging CAD areas should FDA by aware of? Do you have comments on the types of testing needed for these emerging technologies compared to the types of testing you have discussed in this meeting?

B. When should devices be evaluated like CAD devices?

Computers are used to assist in the evaluation of radiological images in a wide spectrum of ways. We recognize that these devices have a spectrum of complexity and risk. The level of information needed to support a marketing submission varies accordingly. These devices can be broadly grouped as:

1. Simple functionalities:
   - Image archive and retrieval tools
   - Image display and manipulation tools
   - Image enhancement and filter tools
   - Annotation tools
2. Semi-automated and automated evaluation tools
   - Volume/change estimation tools
   - Constant/uptake measurement tools
   - Organ segmentation tools
   - Virtual colonoscopy
   - Contrast washout estimation
   - Vascular/air pathway segmentation tools
3. Computer prompt tools
   - Non-disease specific
   - Disease specific
4. Computer diagnosis tools (physician-identified candidates)
   - Ultrasound evaluation of breast masses
   - Evaluation of lung nodules
   - Brain perfusion for diagnosis of Alzheimers
5. Computer diagnosis tools (computer identified candidates)
   - Mammography CADe including lesion ranking
   - Lung CADe including probability of malignancy score
6. Computer first reader devices

Topic for discussion

Do you have comments on the levels of testing for the different types of computer-based technologies compared to the testing you have discussed in this meeting?

- ○

VII. Glossary

510(k): short for section 510(k) of Federal Food, Drug, and Cosmetic Act, refers to the process of clearing a device as substantially equivalent to a predicate device for purposes of legally marketing it under the requirements of this section of the act, one of several processes under which a sponsor may apply to market a medical device

AFROC: Alternative free response receiver operator characteristic curve, a data analysis method applied to imaging studies

Approved: A device that has successfully completed the pre-market approval process, see PMA.

AUC: Area under the curve

BCSC: Breast Cancer Surveillance Consortium

CAD: Computer-aided detection or diagnosis

CADe: Computer aided detection

CADx: Computer aided diagnosis

CC: Cranio-caudal mammography view. image of the breast take from above

CDRH: Center for Devices and Radiological Health

Cleared: Cleared is shorthand for a device that has been determined to be substantially equivalent to a previously marketed predicate device. The predicate device may be a device that was marketed prior to the Medical Device Amendments of 1976 or one cleared via a submission under section 510(k) of the act.

Concurrent read: Using the CAD interpretation simultaneously with initial reading of the image

CT: Computed tomography

CTC: Computed tomographic colonography

CT-lung: Computed tomography of the lung

DBM: The method of interpreting multiple-reader-multiple-case studies proposed by Dorfman, Berbaum and Metz

Enrich: To form a study sample by artificially changing the prevalence of cases of interest

FFDM: Full field digital mammography

Field test: A study design in which a device is studied under the conditions of intended use in the population of intended use.

FN: False negative, a negative test result when the underlying condition exists

FP: False positive, a positive test result when the underlying condition doe not exist

FPF: False positive fraction, the proportion of positive test results in the sample of patients without the condition of interest, 1-specificity.

FROC : Free response receiver operator characteristic curve, a data analysis method applied to imaging studies

Ground truth: The presence or absence of the condition of interest. Ground truth is established using a reference standard (gold standard)

JAFROC: Jackknife FROC, a data analysis method applied to imaging studies that combines re-sampling with FROC analysis

MLO: mediolateral oblique view in mammography, an image taken from an oblique or angled view

MQSA: Mammography Quality Standards Act

MRMC: Multi-reader-multi-case, a class of study designs used to test imaging systems and involving several readers interpreting images.

NCI: National Cancer Institute

NPV: Negative predictive value

ODE: FDA/CDRH Office of Device Evaluation

OSB: FDA/CDRH Office of Surveillance and Biometrics

OSEL: FDA/CDRH Office of Science and Engineering Laboratories

PAUC: Partial area under the curve, used in reference to receiver operator characteristic curve studies

PACS: Picture Archiving and Communications System

PMA: Pre-market approval, a process through which devices are shown to be safe and effective for their intended use prior to being legally marketed see section 515 of Federal Food, Drug, and Cosmetic Act. PMA also refers to the actual application for approval submitted by a device sponsor.

PPV: Positive predictive value

Retrospective study: a study using previously collected data.

ROC: Receiver operator characteristic curve,

ROI: Region of interest, a subsection of an image

RR: Relative reduction

Se: Sensitivity, the probability of a positive test result in patients having the condition of interest, TP/(TP+FN)

Sequential read: A reading scenario where an un-aided interpretation of an image is followed by an aided (CAD assisted) interpretation of an image

Sp: Specificity, the probability of a negative test result in patients not having the condition of interest, TN/(TN+FP)

Standalone: A scenario in which a CAD device is used or evaluated without reader interpretation

Stress test: A test of a device on a set of cases that are chosen because they are believed to be difficult to interpret.

TN: True negative, a negative test result on an individual not having the condition of interest

TP: True positive, a positive test result on an individual having the condition of interest.

TPF: True positive fraction, sensitivity

Training set: Set of data or cases used to train or optimize a CAD device

Truthing: The act of determining the ground truth using a reference (gold) standard

VIII. References for which full text is provided to Panel

Abe, H, Heber MacMahon, et al. Computer-Aided Diagnosis in Chest Radiology, Semin Ultrasound CT MRI 25:432-437, 2004

Armato III, S.G., Rachael Y. Roberts, et al., The Lung Image Database Consortium (LIDC): Ensuring the Integrity of Expert-Defined "Truth", Acad Radiol 2007; 14:1455–1463

Armato III S., Image Annotation for Conveying Automated Lung Nodule Detection Results to Radiologists, Acad Radiol 2003; 10:1000-1007
American College of Radiology ACR Appropriateness Criteria®, Solitary Pulmonary Nodule 2005

Astley, S.M., "Computer-based detection and prompting of mammographic abnormalities." British Journal Radiology, 77: S194–S200, 2004

Astley, S.M., "Evaluation of computer-aided detection (CAD) prompting techniques for mammography," British Journal Radiology, 78: S20–S25, 2005

Baker, S.G., P.F. Pinsky, "A Proposed Design and Analysis for Comparing Digital and Analog Mammography: Special Receiver Operating Characteristic Methods for Cancer Screening," *Journal of the American Statistical Association*, 96(454): 421-428, 2001.

Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects receiver operating characteristic analysis. Acad Radiol 2000; 7:341-349.

Berbaum, K.S., Robert T. Caldwell, et al., Does Computer-Aided Diagnosis for Lung Tumors Change Satisfaction of Search in Chest Radiography?, Acad Radiol 2007; 14:1069–1076

Brown, M.S., Jonathan G. Goldin, et al., Computer-aided Lung Nodule Detection in CT: Results of Large-Scale Observer Test; Acad Radiol 2005; 12:681–686

Birdwell, R.L., P. Bandodkar, D.M. Ikeda, "Computer-aided Detection with Screening Mammography in a University Hospital Setting," Radiology 236:451–457, 2005

Breast Cancer Surveillance Consortium: Evaluating Screening Performance in Practice, National Cancer Institute, Division of Cancer Control and Population Sciences, Applied Research Program, April 2004.

Burhenne, WLJ, Wood SA, D'Orsi CJ, et al., "Potential contribution of computer-aided detection to the sensitivity of screening mammography", Radiology 2000; 215:554-562.

Castellino, R.A., "Computer aided detection (CAD): an overview", Cancer Imaging 5, 17-19, 2005.

Costello, P. and Schoepf, U.J., et al, “Computer-Aided Diagnosis as a Second: Spectrum of Findings in CT Studies of the Chest Interpreted as Normal Reader” Chest 2005;128;1517-1523

Chakraborty, D.P. and Winder, L.H.L., "Free-response methodology: Alternate analysis and a new observer-performance experiment," Radiol., vol. 174, pp. 873-881, 1990.

Cupples, T.E., J.E. Cunningham, J.C. Reynolds, “Impact of Computer-Aided Detection in a Regional Screening Mammography Program,” AJR 185:944–950, 2005

Dachman, A.H. and Zalis, M.E. , “Quality and consistency in CT colonography and research reporting”, Radiology 2004; 230:319–323.

Dean, J.C. and Ilvento, C.C. , “Improved Cancer Detection Using Computer-Aided Detection with Diagnostic and Screening Mammography: Prospective Study of 104 Cancers,” AJR 187:20–28, 2006

Dodd, L.E., Wagner, R.F., et al., “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 2004; 11:462–475.

D’Orsi, C.J., “Computer-aided detection: There is no free lunch”, Radiology 2001; 221:585-586.

D’Orsi, C.J. and Kopans, D.B., “Mammographic Feature Analysis“, Seminars in Roentgenology, vol. 28, no. 3 (July) 1993: p.204-230.

Elmore, J.G. and Carney, P.A., “Computer-Aided Detection of Breast Cancer: Has Promise Outstripped Performance?”, Journal of the National Cancer Institute, vol. 96, no. 3, February 4, 2004.

Fawcett, T. "An introduction to ROC analysis", Pattern Recognition Letters, 27, 861-874, 2006.

“FDA Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests”, 2007, (http://www.fda.gov/cdrh/osb/guidance/1620.pdf)

Fenton, J.J. et al., “Influence of Computer-Aided Detection on Performance of Screening Mammography,” N Engl J Med 356:1399-409, 2007.

Freeman, D.H.,Jr. and Holford, T.R., “Summary Rates,” *Biometrics* 36(2): 195-205, 1980.

Freer, T.W. and Ulissey, M.J., “Screening Mammography with Computer-aided Detection: Prospective Study of 12,860 Patients in a Community Breast Center,” Radiology 220:781–786, 2001.

Goergen, S., et al. “Characteristics of Breast Cancers Missed by Screening Radiologists”, Radiology 204:131–135, 1997.

Gromet, M., "Comparison of Computer-Aided Detection to Double Reading of Screening Mammograms: Review of 231,221 Mammograms", AJR: 190, April 2008

Gur, D., et al., "Changes in Breast Cancer Detection and Mammography Recall Rates After the Introduction of a Computer-Aided Detection System," J Natl Cancer Inst 96:185–90, 2004.

Hadjiiski, L., Sahiner, B., Chan, H.P., "Advances in computer-aided diagnosis for breast cancer", Current Opinion in Obstetrics and Gynecology 18: 64-70, 2006.

Halligan, S. and Altman, D.G., et al.," CT Colonography in the Detection of Colorectal Polyps and Cancer: Systematic Review, Meta-Analysis, and Proposed Minimum Data Set for Study Level Reporting", Radiology 2005; 237:893–904

Helvie M.A. "Improving Mammographic Interpretation: Double Reading and Computer-Aided Diagnosis." Radiol-Clin-N-Am 45:801, 2007.

Helvie, M.A., et al. "Sensitivity of Noncommercial Computer-aided Detection System for Mammographic Breast Cancer Detection: Pilot Clinical Trial," Radiology 231:208–214, 2004.

Hendrick, R. et al. "Community-Based Mammography Practice: Services, Charges, and Interpretation Methods", AJR: 184, February 2005.

Hillis SL, Berbaum KS. Power estimation for the Dorfman-Berbaum-Metz method. Acad Radiol 2004; 11; 1260–1273.

Hillis, S.L., N.A. Obuchowski, K.M. Schartz, K.S. Berbaum, "A comparison of the Dorfman–Berbaum–Metz and Obuchowski–Rockette methods for receiver operating characteristic (ROC) data," Statist. Med. 24:1579–1607, 2005.

Kallergi, M., Carney, G.M., and Gaviria, J. "Evaluating the performance of detection algorithms in digital mammography", Med Phys 26(2), February, 1999.

Ko, J.M., M.J. Nicholas, J.B. Mendel, P.J. Slanetz, "Prospective Assessment of Computer-Aided Detection in
Interpretation of Screening Mammography," AJR 187:1483–1491, 2006

Kondratovich, M.V., "Comparing Two Medical Tests When Results of Reference Standard Are Unavailable for Those Negative via Both Tests," Journal of Biopharmaceutical Statistics, 18: 145–166, 2008.

Li, F. and Engelmann, R., et al., "Lung Cancers Missed on Chest Radiographs: Results Obtained with a Commercial Computer-aided Detection Program", Radiology, 2008, vol 246, page 273

Macari, M. and Bini, E.J., "CT Colonography: Where Have We Been and Where Are We Going?", Radiology 2005; 237:819–833

Morton, M.J., Whaley, D.H., Brandt, K.R., and Amrami, K.K., “Screening Mammograms: Interpretation with Computer-aided Detection—Prospective Evaluation,” Radiology 239:375–383, 2006.

Metz, C.E., "ROC analysis in medical imaging: a tutorial review of the literature", Radiol Phys Technol 1:2-12, 2008.

Mulhall, B.P., Ganesh R., and Veerappan, M., et al., MPH Meta-Analysis: Computed Tomographic Colonography: Ann Intern Med. 2005;142:635-650.

Nishikawa, R.M., Kallergi, M., and Orton, C.G., “Computer-aided Detection, in Its Present Form, is Not an Effective Aid for Screening”, Med Phys 26(2), February, 1999.

Nodine, C. et al, “Time Course of Perception and Decision Making During Mammographic Interpretation”, AJR: 179, October 2002.

Obuchowski NA, Lieber ML, Powell KA, “Data analysis for detection and localization of multiple abnormalities with application to mammography,” Acad. Radiol. 2000; 7: 516-525.

Obuchowski NA, “Multireader receiver operating characteristic studies: a comparison of study designs”, Acad Radiol 1995 ;2:709-716.

Park, S.H., Ha, H.K., et al., “False-Negative Results at Multi–Detector Row CT Colonography: Multivariate Analysis of Causes for Missed Lesions”, Radiology 2005; 235:495–502.

Pickhardt, P.J., et al., “CT Colonography (Virtual Colonoscopy): A Practical Approach for Population Screening”, Radiol Clin N Am 45 (2007) 361–375.

Pickhardt, Perry J., “Screening CT Colonography: How I Do It”, AJR 2007; 189:290–298.

Reeves, A.P. and Kostis, W.J., “Computer-Aided Diagnosis for Lung Cancer”, Radiologic Clinics of North America, vol. 38, no. 3, 2000, pg 497.

Reeves, A.P. and Biancardi, A.M., et al., “The Lung Image Database Consortium (LIDC): A Comparison of Different Size Metrics for Pulmonary Nodule Measurements”, Acad Radiol 2007; 14:1475–1485.

Rockey, D. “Colon Cancer Screening, Polyp Size, and CT Colonography: Making Sense of It All?”, Gastroenterology 131, 2006-2009, 2006.

Roehrig, J. “The Manufacturer’s Perspective”, BJR, 78(2005), 341-345.

Rosenberg, R., et al, “Performance Benchmarks for Screening Mammography”, Radiology, vol. 241, no. 1, October 2006.

Rusinek, H. and Naidich, D.P., et al., “Pulmonary Nodule Detection: Low-Dose versus Conventional CT”, Radiology 1998; 209:243-249.

Sacks W., “Estimating the effect of computer-aided detection on the sensitivity of screening mammography,”Radiology, Volume 226, No 2, 597-599, 2003.

Sickles, E.A., Wolverton, D.E., and Dee, K.E., “Performance Parameters for Screening and Diagnostic Mammography: Specialist and General Radiologists”, Radiology, vol. 224, no. 3, September 2002.

Sickles, E., “Mammographic Features of 300 Consecutive Nonpalpable Breast Cancers”, AJR 146: 661-663, April 1986.

Sickles, E., et al, “Performance Benchmarks for Diagnostic Mammography”, Radiology, vol. 235, no. 3, June 2005.

Siegel, E., et al, “Digital Mammography Quality: Image Display”, J Am Coll Radiol 2006; 3: 615-627.

Sluimer, I., et al. “Computer analysis of Computed Tomography Scans of the Lung: A Survey”, IEEE Transactions on Medical Imaging 25, NO. 4, 385-405, 2006.

Taylor, S.A., et al, “Computer-Assisted Reader Software Versus Expert Reviewers for Polyp Detection on CT Colonography,” AJR 2006; 186:696–702.

Taylor, P., et al, “Impact of Computer Aided Detection Prompts on the Sensitivity and Specificity of Screening Mammography”, U.K. Health Technology Assessment, February 2005.

Wagner RF, Metz CE, Campbell G. “Assessment of Medical Imaging Systems and Computer Aids: A Tutorial Review,” Acad-Radiol 14:723-748, 2007.

Yoshida et al., “Within the next five years CT colonography will make conventional colonoscopy obsolete for colon cancer screening”, MedPhys: (2006), 33, No 8, 2679-2682.

Zalis, ME, et al., “CT Colonography Reporting and Data System: A Consensus Proposal”, Radiology 2005, vol. 236, 3-9.

Zalis, M.E., Perumpillichira, J.J., et al, “Tagging-based, Electronically Cleansed CT Colonography: Evaluation of Patient Comfort and Image Readability”, Radiology 2006, vol. 239, no.1,149-159.

http://www.anaesthetist.com/mnm/stats/roc/Findex.htm, an interactive web site.

IX. Endnotes (Bibliography)

[1] BJ Erickson & B Bartholmai, *Computer-Aided Detection and Diagnosis at the Start of the Third Millennium*, 15(2) Journal of Digital Imaging 59 (2002).
[2] American College of Radiology (ACR) Breast Imaging Reporting and Data System Atlas (BI-RADS® Atlas), Fourth Edition, Reston, Va (2003).
[3] Id.
[4] See 21 CFR 900.12 (Mammography Quality Standards), Subsection (c) (Medical records and mammography reports), stating that: "The mammography report shall include the following information: … Overall final assessment of findings, classified in one of the following categories:

- "Negative:" Nothing to comment upon (if the interpreting physician is aware of clinical findings or symptoms, despite the negative assessment, these shall be explained);
- (B) "Benign:" Also a negative assessment;
- (C) "Probably Benign:" Finding(s) has a high probability of being benign;
- (D) "Suspicious:" Finding(s) without all the characteristic morphology of breast cancer but indicating a definite probability of being malignant;
- (E) "Highly suggestive of malignancy:" Finding(s) has a high probability of being malignant;

In cases where no final assessment category can be assigned due to incomplete work-up, "Incomplete: Need additional imaging evaluation" shall be assigned as an assessment and reasons why no assessment can be made shall be stated by the interpreting physician." These correspond to the BI-RADS Final Assessment Categories.
[5] ME Zalis et al., CT Colonography Reporting and Data System: A Consensus Proposal. 236 Radiology 3-9 (2006).
[6] Leonard Berlin, Malpractice Issues in Radiology: Possessing Ordinary Knowledge, 166 AJR 1027 (1996).
[7] Leonard Berlin, Malpractice Issues in Radiology: Errors in Judgment, 166 AJR 1259 (1996).
[8] Leonard Berlin, Malpractice Issues in Radiology: Perceptual Errors, 167 AJR 587, 589 (1997).
[9] E.g., if a CT scan of the abdomen is performed but a portion of the abdomen is excluded from the images.
[10] E.g., if a chest x-ray is performed and not enough x-ray energy is used, the lungs may not be properly visualized.
[11] E.g., when an MRI of the knee is performed, if the knee is improperly flexed there may be distortion of the normal anatomy and improper visualization of internal structures.
[12] Leonard Berlin, Malpractice Issues in Radiology: Picture Archiving and Communications Systems (PACS) and the Loss of Patient Examination Records, 176 AJR 1381 (2001).
[13] EA Krupinski. Computer-aided Detection in Clinical Environment: Benefits and Challenges for Radiologists. 231 Radiology 7 (2004).
[14] Picture from the FDA Display Laboratory.
[15] RS Rana et al., *Independent Evaluation of Computer Classification of Malignant and Benign Calcifications in Full-Field Digital Mammograms*. 14 Acad Radiol 363 (2007).
[16] Image data from USF Digital Database for Screening Mammography (DDSM). See http://marathon.csee.usf.edu/Mammography/Database.html. CAD marks produced by an internal FDA research CAD system (f0037rm92_1)
[17] Castellino, R.A., "Computer aided detection (CAD): an overview", Cancer Imaging (2005) 5, 17–19.
[18] Id. at 296
[19] 21 CFR 860.3(c)(2)
[20] 21 CFR 860.7
[21] The term "pre-amendment device" refers to devices legally marketed in the U.S. before May 28, 1976 (the effective date of the Medical Device Amendments to the FDCA) and which have not been significantly changed or modified since then. These devices were "grandfathered" into the system.
[22] 21 CFR 807.100
[23] See 21 CFR 807.87(g)
[24] See http://www.fda.gov/cdrh/devadvice/314312.html#link_6 (last accessed 1/12/08).
[25] 21 CFR 860.7(d)

[26] 21 CFR 860.7(e).

[27] EA Krupinski et al., A Perceptually Based Method for Enhancing Pulmonary Nodule Recognition. 28(4) Investigative Radiology 289 (1993).

[28] Image data from USF Digital Database for Screening Mammography (DDSM). See http://marathon.csee.usf.edu/Mammography/Database.html. CAD marks produced by an internal FDA research CAD system (f0009rm92_1).

[29] Fryback DG, Thornbury JR. The efficacy of diagnostic imaging. Med Decis Making 1991; 11:88–94.

[30] NCRP Commentary No. 13. An introduction to efficacy in diagnostic radiology and nuclear medicine. Bethesda, Md: National Council on Radiation Protection and Measurements, 1995.

[31] Pepe MS. The statistical evaluation of medical tests for classification and prediction. New York: Oxford University Press, 2003, See Chapter X.

[32] Zhou X-H, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine (Wiley series in probability and statistics). New York: Wiley-Interscience, 2002.

[33] Brealey S and Scally AJ, Bias in plain film reading performance studies, *The British Journal of Radiology*, 74 (2001), 307–316

[34] RF Wagner et al., "Assessment of Medical Imaging Systems and Computer Aids: A Tutorial Review," Academic Radiology 723–748, 2007.

[35] Wagner RF, Metz CE, Campbell G. Assessment of Medical Imaging Systems and Computer Aids: A Tutorial Review. Acad-Radiol 2007; 14:723-748.

[36] Miller DP, O'Shaughnessy KF, Wood SA, Castellino RA. Gold standards and expert panels: a pulmonary nodule case study with challenges and solutions. Proceedings of the SPIE - Medical Imaging 2004; 5372:173-184.

[37] Pepe MS. "The statistical evaluation of medical tests for classification and prediction," New York: Oxford University Press, 2003, See Chapter X.

[38] Obuchowski NA and Zhou XH, Prospective studies of diagnostic test accuracy when disease prevalence is low, *Biostatistics* (2002), 3 (4), 477–492.

[39] Warren Burhenne LJ, Wood SA, D'Orsi CJ, et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. Radiology 2000; 215:554-562.

[40] Petrick N, Haider M, Summers RM, et al. CT colonography with computer-aided detection as a second reader: observer performance study. Radiology 2008; 246:148-156.

[41] Metz CE. Basic principles of ROC analysis. Semin Nucl Med 1978; 8:283–298.

[42] Dodd LE, Wagner RF, Armato SG, et al. 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 (LIDC). Acad Radiol 2004; 11:462–475.

[43] Chakraborty DP. Statistical power in observer performance studies: comparison of the receiver operating characteristic and free-response methods in tasks involving localization. Acad Radiol 2002; 9:147–156.

[44] Metz CE. Evaluation of CAD methods. In: Doi K, MacMahon H, Giger ML, Hoffmann KR, eds. Computer-aided diagnosis in medical imaging. Excerpta Medica International Congress Series, Vol. 1182.

[45] Chakraborty DP, Berbaum KS. Observer studies involving detection and localization: Modeling, analysis, and validation. Med Phys 2004; 31:2313–2330.

[46] E Alberdi, A Povyakalo, L Strigini, and P Ayton, "Effects of Incorrect Computer-aided Detection (CAD) Output on Human Decision-making in Mammography," Acad. Radiol. 2004; 11:909–918.

[47] Hillis SL, Berbaum KS. Power estimation for the Dorfman-Berbaum-Metz method. Acad Radiol 2004; 11; 1260–1273.

[48] Zhou X-H, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine (Wiley series in probability and statistics). New York: Wiley-Interscience, 2002.

[49] Dorfman DD, Berbaum KS, Metz CE. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. Invest-Radiol 1992; 27:723-731.

[50] Obuchowski NA, Lieber ML, Powell KA, "Data analysis for detection and localization of multiple abnormalities with application to mammography," Acad. Radiol. 2000; 7: 516-525.

[51] Beiden SV, Wagner RF, Campbell G. Components-of-variance models and multiple-bootstrap experiments: an alternative method for random-effects receiver operating characteristic analysis. Acad Radiol 2000; 7:341-349.

[52] Liang KY and Zeger SL, 1986, "Longitudinal data analysis using generalized linear models," Biometrika, 73, 13-22.
[53] Milliken and Johnson, Analysis of Messy Data; Volume I: Design Experiments, 1984, Van Nostrand, Reinhold.
[54] Freeman and Holford, Summary Rates, Biometrics, 1980, 195-205
[55] PMA P020014, Conceptus Essure, Summary of Safety and Effectiveness Data, http://www.fda.gov/cdrh/pdf2/p020014b.pdf.
[56] Schatzkin A, Connor RJ, Taylor PF, and Bunnag B, Comparing new and old screening tests when a reference procedure cannot be performed on all screenees, American Journal of Epidemiology, 1987, 25, 672-678.

[57] Kondratovich. Marina V. (2008), Comparing Two Medical Tests When Results of Reference Standard Are Unavailable for Those Negative via Both Tests, Journal of Biopharmaceutical Statistics, 18:1, 145-166.

[58] Use of Mammograms Among Women Aged >40 Years --- United States, 2000—2005, 56(3) MMWR 49-51, January 26, 2007.
59 RA Smith et al., Cancer Screening in the United States, 2007: A Review of Current Guidelines, Practices, and Prospects, 57 CA Cancer J Clin 90-104, 2007.
60 See http://www.fda.gov/CDRH/MAMMOGRAPHY/scorecard-statistics.html.
61 See http://breastscreening.cancer.gov/data/variables/freq_tables_pct.html#indication
62 I Andersson, Breast Cancer Screening with Mammography, 2 Breast Care 4-5, 2007.
63 Id.
64 See e.g., K Armstrong et al., Screening Mammography in Women 40 to 49 Years of Age: A Systematic Review for the American College of Physicians, 146 Ann Intern Med 516-526, 2007. (stating that: "Current evidence indicates that women 40 to 49 years of age who undergo routine mammography screening will decrease their risk for death due to breast cancer but will increase their risks for undergoing unnecessary procedures, breast cancer–related anxiety, discomfort at the time of screening, and exposure to low-dose radiation. Because the incidence of breast cancer and the effectiveness of mammography are lower among women in their 40s than among women 50 years of age or older, mammography screening results in less absolute benefit and greater absolute risk for women 40 to 49 years of age than for women 50 years of age or older. The proportion of women 50 years of age or older whose risks for mammography outweigh the benefits is widely accepted to be clinically insignificant. However, the evidence suggests that this proportion is higher and may be clinically significant for women 40 to 49 years of age. Given this difference, a woman 40 to 49 years of age who had a lower-than-average risk for breast cancer and higher-than-average concerns about false positive results might reasonably delay screening. Measuring risks and benefits accurately enough to identify these women remains a challenge."); See also I Andersson, Breast Cancer Screening with Mammography, 2 Breast Care 4-5, 2007 (stating that "screening mammography in [the 40-50 year-old] age group has remained a topic of debate because of the difficulty in determining the benefit of mammography in this age group. … Because screening mammography is also associated with potential harms, a discussion of risks (biopsies, surgery, radiation exposure, false-positive results, and false reassurance), benefits (early detection of breast cancer), and patient preferences should be the basis for screening decisions.")
65 American Cancer Society Guidelines Still Advise Yearly Mammography For Women Aged 40 To 49 Years, 57 CA Cancer J Clin 187-188, 2007.
66 ACR Practice Guideline For The Performance Of Screening Mammography (effective 10/01/04).
67 American College of Obstetricians and Gynecologists, Committee Opinion No. 356: Routine Cancer Screening, 108 Obstetrics & Gynecology 1611-1613, December 2006; See
68 A Qassem et al., Screening Mammography for Women 40 to 49 Years of Age: A Clinical Practice Guideline from the American College of Physicians, 146 Ann Intern 511-515, 2007.
69 See Summaries for Patients, *Screening for Breast Cancer: Recommendations from the U.S. Preventive Services Task Force*, 137 Ann Int Med I-47, 2002 (stating that: "Experts do not all agree about whether mammography decreases breast cancer death rates. … The disagreement is strongest for mammography in 40- to 49-year-old women. … The potential harms of mammography include anxiety, procedures, and costs
that result from mammograms that suggest cancer when there is none (false positives).")
70 See, Recommendations for Clinical Preventive Services, Breast Cancer, at http://www.aafp.org/online/en/home/clinical/exam/a-e.html.
71 See R Harris, *Screening for Breast Cancer: What to Do with the Evidence?*, 75 American Family Physician 1623, 2007 (stating that: "To summarize the latest evidence, … for women in their 40s: screening probably decreases breast cancer mortality, but only to a small degree. … But none of the studies tells us whether these technologies save more lives or are worth the additional false-positive results, overdiagnosis, and cost."

[72] RD Rosenberg et al., *Performance Benchmarks for Screening Mammography*, 241 Radiology 55-66, 2006.
[73] JG Elmore et al., *Ten-Year Risk Of False Positive Screening Mammograms And Clinical Breast Examinations*, 338 NEJM 1089-1096, 1998.
[74] See R Harris, Screening for Breast Cancer: What to Do with the Evidence?, 75 American Family Physician 1623, 2007.
[75] Id.
[76] S Zackrisson et al., *Rate of over-diagnosis of breast cancer 15 years after end of Malmö mammographic screening trial: follow-up study*, 332 BMJ 689-692, 2006.
[77] See e.g., EA Sickles, *Mammographic Features of 300 Consecutive Non-palpable Breast Cancers*. 146 AJR 661 (1986)
[78] See BT Nicholson et al., *Accuracy of Assigned BI-RADS Breast Density Category Definitions*, 13 Acad Radiol 1143–1149, 2006.
[79] See e.g., RE Bird et al., *Analysis of Cancers Missed at Screening Mammography*. 184 Radiology 613 (1992). VP Jackson et al., *Imaging of the Radiographically Dense Breast*, 188 Radiology 297 (1993); MT Mandelson et al., *Breast Density as a Predictor of Mammographic Detection: Comparison of Interval- and Screen-Detected Cancers*. 92(13) JNCI 1081 (2000).
[80] See e.g., JA Tice et al., *Mammographic Breast Density And The Gail Model For Breast Cancer Risk Prediction In A Screening Population.* 94 Breast Cancer Research and Treatment 115 (2005).
[81] ED Pisano et al., Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening. 353(17) NEJM 1773 (2005).
[82] See http://breastscreening.cancer.gov/data/variables/freq_tables_pct.html#density
[83] See e.g., EA Sickles, *Mammographic Features of 300 Consecutive Non-palpable Breast Cancers*. 146 AJR 661 (1986); AM Knutzen, *Likelihood Of Malignant Disease For Various Categories Of Mammographically Detected, Nonpalpable Breast Lesions*. 68 Mayo Clin Proc 454 (1993); LJW Burhenne et al., *Potential Contribution of Computer-aided Detection to the Sensitivity of Screening Mammography*. 215 Radiology 554 (2000); RL Birdwell et al., *Computer-aided Detection with Screening Mammography in a University Hospital Setting*. 236 Radiology 451 (2005).
[84] See e.g., ZS Cosar et al., *Concordance of mammographic classifications of microcalcifications in breast cancer diagnosisB Utility of the Breast Imaging Reporting and Data System (fourth edition)*. 29 J Clin Imaging 389 (2005); L Lieberman et al., *The Breast Imaging Reporting and Data System: Positive PredictiveValue of MammographicFeatures and Final Assessment Categories*. 171 AJR 35 (1998); B de Lafontan et al., *Isolated Clustered Microcalcifications: Diagnostic Value of Mammography Series of400 Cases with Surgical Verification*. 190 Radiology 479 (1994).
[85] See e.g., KA Skinner, *Palpable Breast Cancers Are Inherently Different From Nonpalpable Breast Cancers*. 8 Annals of Surgical Oncology 705 (2001); P Perdue. *Early Detection Of Breast Carcinoma: A Comparison Of Palpable And Nonpalpable Lesions*. 111(6) Surgery 656 (1992); MJ Kim et al., *Comparison between Palpable and Nonpalpable Breast Cancers: Mammographic and Pathological Findings*. 41(1) J Korean Radiol Soc 181 (1999).
[86] AG Haus & MJ Yaffee, *Screen-Film and Digital Mammography Image Quality and Radiation Dose Considerations*, 38 Radiologic Clinics of North America 871-898, 2000.
[87] MB Williams et al., *Image Quality in Digital Mammography: Image Acquisition*, 3 JACR 589-608, 2006.
[88] Id.
[89] E Siegel et al., Digital Mammography Image Quality: Image Display, 3 JACR 615-627, 2006.
[90] Id.
[91] EA Krupinski et al., A Perceptually Based Method for Enhancing Pulmonary Nodule Recognition. 28(4) Investigative Radiology 289 (1993).
[92] E Siegel et al., Digital Mammography Image Quality: Image Display, 3 JACR 615-627, 2006.

[93] Id. (citing B Zheng et al., Classification of Breast Abnormalities Under Different Detection Cueing Environments, Proc SPIE 5372:211-7, 2004).

[94] See http://marathon.csee.usf.edu/Mammography/DDSM/thumbnails/normals/normal_01/overview.html

[95] MB Williams et al., *Image Quality in Digital Mammography: Image Acquisition*, 3 JACR 589-608, 2006.

[96] See 21 CFR Part 900.

[97] 21 CFR 900.1.

[98] 21 CFR 900.12(c).

[99] See http://www.acr.org/SecondaryMainMenuCategories/quality_safety/BIRADSAtlas.aspx

[100] See 21 CFR 900.12(c)(iv)

[101] See ACR Practice Guideline For The Performance Of Screening Mammography.

[102] See ACR Practice Guideline For The Performance Of Diagnostic Mammography.

[103] BM Geller et al., *The use of Additional Imaging Increased Specificity and Decreased Sensitivity in Screening Mammography*, 58 Journal of Clinical Epidemiology 942–950, 2005.

[104] See ACR BI-RADS Atlas, Fourth Edition (2003), Mammography, Masses, page 9.

[105] See ACR BI-RADS Atlas, Fourth Edition (2003), Mammography, Guidance Chapter, page 253.

[106] See ACR BI-RADS Atlas, Fourth Edition (2003), Mammography, Calcifications, page 118.

[107] ACR BI-RADS Mammography, Introduction, page 1.

[108] ACR BI-RADS Mammography, Guidance Chapter, page 254.

[109] http://breastscreening.cancer.gov/about/

[110] See, RD Rosenberg et al., *Performance Benchmarks for Screening Mammography*, 241 Radiology 55-66, 2006; EA Sickles et al., *Performance Benchmarks for Diagnostic Mammography*, 235 Radiology 775–790, 2005.

[111] See M Helvie, *Improving Mammographic Interpretation: Double Reading and Computer-Aided Diagnosis*, 45 Radiologic Clinics of North America 801–811, 2007. (citing CA Beam, Reader Strategies: Variability and Error Methodology, Findings, and Health Policy Implications from a Study of the U.S. Population of Mammographers. (In: DP Chakraborty & EA Krupinski (editors), Medical imaging 2002: Image Perception, Observer Performance, and Technology Assessment, San Diego, CA, SPIE: 157–68, 2002; and CA Beam et al., *Association of Volume and Volume-Independent Factors with Accuracy in Screening Mammogram Interpretation*, 95 JNCI 282–90, 2003).

[112] See http://breastscreening.cancer.gov/data/benchmarks/screening/tableSensSpec.html

[113] JM Lewin et al., *Digital Mammography*, 42 Radiologic Clinics of North America 871 – 884, 2004.

[114] RE Bird et al., Analysis of Cancers Missed at Screening Mammography, 184 Radiology 613–617, 1992; SK Goergen, et al., *Characteristics of Breast Carcinomas Missed by Screening Radiologists*, 204 Radiology 131–135, 1997; TM Kolb et al., *Comparison of the Performance ofSscreening Mammography, Physical Examination, and Breast US and Evaluation of Factors that Influence Them: An Analysis of 27,825 Patient Evaluations*, 225 Radiology 165-175, 2002.

[115] BC Yankaskas et al., *Association of Recall Rates with Sensitivity and Positive Predictive Values of Screening Mammography*, 177 AJR 543–549, 2001 (finding that: "recall rates of the average practice ranged from 1.9% to 13.4%. Sensitivity rose from a mean of 65% in the lowest recall rates to 80.2% at the highest level of recall rates. The positive predictive value of screening decreased from 7.2% in the lowest level of recall to 3.3% in the highest. As recall rates increased, sensitivity increased very little beyond a recall rate of 4.8%, and positive predictive value began decreasing significantly at a recall rate of 5.9%."); D Gur et al., *Recall and Detection Rates in Screening Mammography. A Review of Clinical Experience—Implications for Practice Guidelines*, 100 Cancer 1590–1594, 2004 (finding "a statistically significant correlation ($P < 0.05$) between recall and detection rates [in a] group of 10 experienced radiologists. … This increase in detection rate was found to persist over the range of observed recall rates and extended beyond the currently recommended practice guideline of 10%."); JDM Otten et al., *Effect of Recall Rate on Earlier Screen Detection of Breast Cancers Based on the Dutch Performance Indicators*, 97 J Natl Cancer Inst 748-754, 2005 (finding that: "Breast cancer can be detected earlier by lowering the threshold for recall, especially for recall rates of 1%–4%. With further recall rate increases, cancer detection levels off with a disproportionate increase of false-positive rates.").

[116] CF Nodine et al., *Time Course of Perception and Decision Making During Mammographic Interpretation*, 179 AJR 917–23, 2002; EA Krupinski, *Visual Search of Mammographic Images: Influence of Lesion Subtlety*, 12 Acad Radiol 965–969, 2005; C Mello-Thoms et al., *Effects of Lesion Conspicuity on Visual Search in Mammogram Reading*, 12 Acad Radiol 830–40, 2005.

[117] See LL Humphrey et al., *Breast Cancer Screening: A Summary of the Evidence for the U.S. Preventive Services Task Force*, 137 Ann Intern Med 347-360, 2002.

[118] BC Yankaskas et al., *Association of Recall Rates with Sensitivity and Positive Predictive Values of Screening Mammography*, 177 AJR 543–549, 2001 (finding that: "recall rates of the average practice ranged from 1.9% to 13.4%. Sensitivity rose from a mean of 65% in the lowest recall rates to 80.2% at the highest level of recall rates. The positive predictive value of screening decreased from 7.2% in the lowest level of recall to 3.3% in the highest. As recall rates increased, sensitivity increased very little beyond a recall rate of 4.8%, and positive predictive value began decreasing significantly at a recall rate of 5.9%."); D Gur et al., *Recall and Detection Rates in Screening Mammography. A Review of Clinical Experience—Implications for Practice Guidelines*, 100 Cancer 1590–1594, 2004 (finding "a statistically significant correlation ($P < 0.05$) between recall and detection rates [in a] group of 10 experienced radiologists. … This increase in detection rate was found to persist over the range of observed recall rates and extended beyond the currently recommended practice guideline of 10%."); JDM Otten et al., *Effect of Recall Rate on Earlier Screen Detection of Breast Cancers Based on the Dutch Performance Indicators*, 97 J Natl Cancer Inst 748-754, 2005 (finding that: "Breast cancer can be detected earlier by lowering the threshold for recall, especially for recall rates of 1%–4%. With further recall rate increases, cancer detection levels off with a disproportionate increase of false-positive rates.").

[119] CF Nodine et al., *Time Course of Perception and Decision Making During Mammographic Interpretation*, 179 AJR 917–23, 2002; EA Krupinski, *Visual Search of Mammographic Images: Influence of Lesion Subtlety*, 12 Acad Radiol 965–969, 2005; C Mello-Thoms et al., *Effects of Lesion Conspicuity on Visual Search in Mammogram Reading*, 12 Acad Radiol 830–40, 2005.

[120] CF Nodine et al., *Time Course of Perception and Decision Making During Mammographic Interpretation*, 179 AJR 917–23, 2002.

[121] M Helvie, *Improving Mammographic Interpretation: Double Reading and Computer-Aided Diagnosis*, 45 Radiologic Clinics of North America 801–811, 2007.

[122] Id.

[123] SM Durfee et al., *Sonographic Evaluation of Clinically Palpable Breast Cancers Invisible on Mammography*, 6 The Breast Journal 247-251, 2000.

[124] See e.g., RE Bird et al. , Analysis of Cancers Missed at Screening Mammography, 184 Radiology 613-617, 1992; JAAM van Dijck et al., The Current Detectability of Breast Cancer in a Mammographic Screening Program, 72 Cancer 1933-1938, 1993.

[125] JA Harvey et al., *Previous Mammograms in Patients with Impalpable Breast Carcinoma: Retrospective vs. Blinded Interpretation*, 161 AJR 1167-1172, 1993.

[126] RE Bird et al. , Analysis of Cancers Missed at Screening Mammography, 184 Radiology 613-617, 1992.

[127] C Mello-Thomas, *Perceptual Analysis of Cancers Missed in Mammography Screening*, Proceedings of SPIE -- Volume 5034, Medical Imaging 2003: Image Perception, Observer Performance, and Technology Assessment, May 2003, pp. 253-256; C Mello-Thomas et al., *What Attracts the Eye to the Location of Missed and Reported Breast Cancers?*, Proceedings Of The 2002 Symposium On Eye Tracking Research & Applications, New Orleans, Louisiana, SESSION: Eye Movement Analysis & Visual Search, 111 – 117, 2002.

[128] C Mello-Thomas et al., *Effects of Lesion Conspicuity on Visual Search in Mammogram Reading*, 12 Acad Radiol 830–840, 2005 (finding that: "A total of 66% of the malignant masses in the current mammograms and 57% in the priors attracted some amount of visual attention. From these, 71% yielded a report on the current mammograms, but only 40% on the priors. In the cases where the observer saw the malignant mass, they did so within 2 seconds of image display, regardless of whether the mammogram was current or prior.").

---

[129] RG Blanks et al., *Observer Variability In Cancer Detection During Routine Repeat (Incident) Mammographic Screening In A Study Of Two Versus One View Mammography*, 6 J Med Screen 152–158, 1999 (finding that: "for the cancers that were not recalled, at least 37% were detected … but incorrectly classified as benign and not recalled. In practice, considerably more than 37% might have been detected, but many readers omitted to mark any features at all if they did not recall (only providing an obligatory do not recall). An analysis using only those forms where the reader provided a high level of detail suggests that the proportion of cancers detected on the film but incorrectly considered to be not suggestive of malignancy is much higher and could be the majority of non-recalled cancers.").

[130] RG Blanks et al., *Observer Variability In Cancer Detection During Routine Repeat (Incident) Mammographic Screening In A Study Of Two Versus One View Mammography*, 6 J Med Screen 152–158, 1999.

[131] See e.g., RL Birdwell et al., *Mammographic Characteristics of 115 Missed Cancers Later Detected with Screening Mammography and the Potential Utility of Computer-aided Detection*, 219 Radiology 192-202, 2001; D Reintgen et al., *The Anatomy of Missed Breast Cancers*, 2 Surg Oncol 65-75, 1993;

[132] PA Carney et al., *Individual and Combined Effects of Age, Breast Density, and Hormone Replacement Therapy Use on the Accuracy of Screening Mammography*, 138 Ann Intern Med 168-175, 2003 (finding that "Adjusted sensitivity ranged from 62.9% in women with extremely dense breasts to 87.0% in women with almost entirely fatty breasts. … Adjusted specificity increased from 89.1% in women with extremely dense breasts to 96.9% in women with almost entirely fatty breasts.")

[133] D Georgian-Smith, et al., Blinded Comparison of Computer-Aided Detection with Human Second Reading in Screening Mammography, 189 AJR 1135–1141, 2007; M Helvie, *Improving Mammographic Interpretation: Double Reading and Computer-Aided Diagnosis*, 45 Radiologic Clinics of North America 801–811, 2007.

[134] M Helvie, *Improving Mammographic Interpretation: Double Reading and Computer-Aided Diagnosis*, 45 Radiologic Clinics of North America 801–811, 2007 (see Table 1).

[135] M Helvie, *Improving Mammographic Interpretation: Double Reading and Computer-Aided Diagnosis*, 45 Radiologic Clinics of North America 801–811, 2007.

[136] Id.

[137] F Shtern, *Digital mammography and related technologies: a perspective from the National Cancer Institute*, 183 Radiology 629-30, 1992.

[138] See http://www.fda.gov/ohrms/dockets/ac/98/transcpt/3418t1.pdf.

[139]

[140]

[141] E Thurfjell et al., *Sensitivity and Specificity of Computer-Assisted Breast Cancer Detection in Mammography Screening*, 39 Acta Radiologica 384–88, 1998.

[142] S Ciatto et al., *Comparison of Standard Reading and Computer-Aided Detection (CAD) on a National Proficiency Test of Screening Mammography*, 45 Eur J Radiol 135–138, 2003.

[143] PM Taylor PM et al., An evaluation of the impact of computerbased prompts on screen readers' interpretation of mammograms, 77 Br J Radiol 21-27, 2004.

[144] Alberdi E, Povyakalo A, Strigini L, et al., *Effects of Incorrect Computer-Aided Detection (CAD) Output on Human Decision-Making in Mammography*, 11 Acad Radiol 909–918, 2004.

[145] SV Destounis et al., *Can Computer-Aided Detection with Double Reading of Screening Mammograms Help Decrease the False-Negative Rate? Initial Experience*, 232 Radiology 578–584, 2004.

[146] TW Freer & MJ Ulissey, *Screening Mammography with Computer-Aided Detection: Prospective Study of 12,860 Patients in a Community Breast Center*, 220 Radiology 781–786, 2001.

[147] D Gur et al., *Changes in Breast Cancer Detection and Mammography Recall Rates After the Introduction of a Computer- Aided Detection System*, 96 J Natl Cancer Inst 185–190, 2004.

[148] RL Birdwell et al., *Computer-aided Detection with Screening Mammography in a University Hospital Setting*, 236 Radiology 451–457, 2005.

[149] TE Cupples et al., *Impact of Computer-Aided Detection in a Regional Screening Mammography Program*, 185 AJR 944–950, 2005.

[150] MJ Morton et al., Screening Mammograms: Interpretation with Computer-Aided Detection–Prospective Evaluation, 239 Radiology 375–383, 2006.

[151] JC Dean & CC Ilvento, *Improved Cancer Detection using Computer-Aided Detection with Diagnostic and Screening Mammography: Prospective Study of 104 Cancers*, 187 AJR 20–28, 2006.
[152] JM Ko et al., *Prospective Assessment of Computer-Aided Detection in Interpretation of Screening Mammography*, 187 AJR 1483–1491, 2006.
[153] JJ Fenton et al., *Influence of Computer-Aided Detection on Performance of Screening Mammography*, 356 NEJM 1399-1409, 2007.
[154] D Gerogian-Smith et al., *Blinded Comparison of Computer-Aided Detection with Human Second Reading in Screening Mammography*, 189 AJR 1135–1141, 2007.
[155] RL Ellis et al., *Evaluation of Computer-aided Detection Systems in the Detection of Small Invasive Breast Carcinoma*, 245 Radiology 88-94, 2007.
[156] SH Taplin et al., *Testing the Effect of Computer- Assisted Detection on Interpretive Performance in Screening Mammography*, 187 AJR 1475–1482, 2006.
[157] Gromet, M., Comparison of Computer-Aided Detection to Double Reading of Screening Mammograms: Review of 231,221 Mammograms, AJR:190, April 2008
[158] See http://breastscreening.cancer.gov/data/benchmarks/screening/tableSensSpec.html
[159] ED Pisano et al., *Diagnostic Performance of Digital versus Film Mammography for Breast-Cancer Screening*. 353(17) NEJM 1773 (2005).
[160] RF Wagner et al., *Assessment of Medical Imaging and Computer-assist Systems: Lessons from Recent Experience*, 9 Acad Radiol 1264–1277, 2002.
[161] See e.g., EA Sickles, *Mammographic Features of 300 Consecutive Nonpalpable Breast Cancers*, 146 AJR 661-663, 1996; RE Bird et al., *Analysis of Cancers Missed at Screening Mammography*, 184 Radiology 613-617, 1992; SK Goergen et al., *Characteristics of Breast Carcinomas Missed by Screening Radiologists*, 204 Radiology 131-135, 1997; JA Harvey et al., *Previous Mammograms in Patients with Impalpable Breast Carcinoma: Retrospective vs. Blinded Interpretation*, 161 AJR 1167-1172, 1993.
[162] *Comparative Specifications of Full Field Digital Mammography Systems* www.hmso.gov.uk
[163] Martin Yaffe, Aili K. Bloomquist, Gordon Mawdsley, Etta D. Pisano, R. Edward Hendrick, Laurie L. Fajardo, John M. Boone, Kalpana Kanal, Mahadevappa Mahesh, Richard C. Fleiscman, Joseph Och, Mark Willams, Daniel J. Beideck, Andew D.A. Maidmont, Quality Control for digital mammography: Part II recommendations from the ACRIN DMIST trial, Med. Phys. 33 (3), 737-752 (2006).
[164] AJ Maltino et al., Practice Patterns of Radiologists and Nonradiologists in Utilization of Noninvasive Diagnostic Imaging among the Medicare Population 1993–1999, 228 Radiology 795-801, 2003.
[165] See e.g., ACR Appropriateness Criteria, revealing 43 separate categories of indications that relate to chest/lung imaging and each category may have multiple variants—e.g., the "Routine Chest Radiograph has 7 variants.
[166] See ACR Practice Guideline for the Performance of Pediatric and Adult Thoracic Computed Tomography (CT)
[167] See ACR Practice Guideline For The Performance Of Pediatric And Adult Thoracic Computed Tomography (CT).
[168] See ACR Practice Guideline for the Performance of High-Resolution Computed Tomography (HRCT) of the Lungs in Adults.
[169] See http://www.cdc.gov/cancer/lung/statistics/
[170] Id.
[171] See e.g., J Shiraishi et al., *Computer-aided Diagnosis for the Detection and Classification of Lung Cancers on Chest Radiographs: ROC Analysis of Radiologists' Performance*, 13 Academic Radiology 995–1003, 2006; CI Henschke et al., Survival of Patients with Stage I Lung Cancer Detected on CT Screening, 355 NEJM 1763-1771, 2006.
[172] See e.g., HG Welch et al., *Overstating the Evidence for Lung Cancer Screening: The International Early Lung Cancer Action Program (I-ELCAP) Study*, 167 Archives of Internal Medicine 2289-2295, 2007; CI Henschke et al., *International Early Lung Cancer Early Action Program Investigators. Survival of Patients with Stage I Lung Cancer Detected on CT Screening*, 355 New England Journal of Medicine 1763-1771, 2006.
[173] See American College of Radiology ACR Appropriateness Criteria, Solitary Pulomary Nodule.

[174] H McMahon et al., *Guidelines for Management of Small Pulmonary Nodules Detected on CT Scans: A Statement from the Fleischner Society*, 237 Radiology 395-400, 2005.
[175] Id.
[176] DE Midthun, Solitary Pulmonary Nodule: Time to Think Small, 6 Curr Opin Pulm Med 364–370, 2000.
[177] See American College of Radiology ACR Appropriateness Criteria, Solitary Pulomary Nodule.
[178] Lillington G. *Disease of the Month*. Mosby-Year Book, Inc; 1991.
[179] See American College of Radiology ACR Appropriateness Criteria, Solitary Pulomary Nodule.
[180] See LGBA Quekel et al., Miss Rate of Lung Cancer on the Chest Radiograph in Clinical Practice, 115 Chest 720-724, 1999.
[181] See e.g., See LGBA Quekel et al., Miss Rate of Lung Cancer on the Chest Radiograph in Clinical Practice, 115 Chest 720-724, 1999; F Li et al., Lung Cancers Missed on Chest Radiographs: Results Obtained with a Commercial Computer-aided Detection Program, 246 Radiology 273-280, 2008.
[182] CL Henschke et al., *Early Lung Cancer Action Project: Overall Design and Findings from Baseline Screening*. 354 Lancet 99–105. 1999, SJ Swenson et al., *Screening for Lung Cancer with CT: Mayo Clinic Experience*. 226 Radiology 756–761, 2003.
[183] H McMahon et al., *Guidelines for Management of Small Pulmonary Nodules Detected on CT Scans: A Statement from the Fleischner Society*, 237 Radiology 395-400, 2005.
[184] Id.
[185] JJ Erasmus et al., *Solitary pulmonary nodules. I. Morphologic Evaluation for Differentiation of Benign and Malignant Lesions*. 20 RadioGraphics 43-58,2000.
[186] SC Gaerte et al., *Fat-Containing Lesions of the Chest*. 22 RadioGraphics S61-S78, 2002.
[187] H McMahon et al., *Guidelines for Management of Small Pulmonary Nodules Detected on CT Scans: A Statement from the Fleischner Society*, 237 Radiology 395-400, 2005.
[188] H Rusinek et al., *Pulmonary Nodule Detection: Low-Dose versus Conventional CT*, 209 Radiology 243-249, 1998.
[189] DP Naidich et al., Variables Affecting Pulmonary Nodule Detection with Computer Tomography: Evaluation with Three-Dimensional Computer Simulation, 8 J Thorac lmaging 291- 299, 1993.
[190] H Rusinek et al., *Pulmonary Nodule Detection: Low-Dose versus Conventional CT*, 209 Radiology 243-249, 1998.
[191] DP Naidich et al., Variables Affecting Pulmonary Nodule Detection with Computer Tomography: Evaluation with Three-Dimensional Computer Simulation, 8 J Thorac lmaging 291- 299, 1993.
[192] See LGBA Quekel et al., Miss Rate of Lung Cancer on the Chest Radiograph in Clinical Practice, 115 Chest 720-724, 1999.
[193] K.S. Berbaum, R.T. Caldwell, K.M. Schartz, B.H. Thompson, and E.A. Franken, Jr, "Does computer-aided diagnosis for lung tumors change satisfaction of search in chest radiography?" Acad. Radiol. 2007; 14:1069–1076
[194] F. Li, R. Engelmann, C.E. Metz, K. Doi, and H. MacMahon, "Lung cancers missed on chest radiographs: Results obtained with a commercial computer-aided detection program," Radiology 2008; 246(1): 273-280
[195] K. Suzuki, J. Shiraishi, H. Abe, H. MacMahon, and K. Doi, "False-positive reduction in computer-aided diagnostic scheme for detecting nodules in chest radiographs by means of massive training artificial neural network," Acad. Radiol. 2005; 12:191–201
[196] . Shiraishi, F. Li, and K. Doi, "Computer-aided diagnosis for improved detection of lung nodules by use of posterior-anterior and lateral chest radiographs," Acad. Radiol. 2007; 14:28–37
[197] A.P. Reeves, and W.J. Kostis, "Computer-aided diagnosis for lung cancer," Radiologic Clinics of North America 2000; 38(3):497-509
[198] H. Arimura, S. Katsuragawa, K. Suzuki, F. Li, J. Shiraishi, S. Sone, and K. Doi, "Computerized scheme for automated detection of lung nodules in low-dose computed tomography images for lung cancer screening," Acad. Radiol. 2004; 11:617–629
[199] S.G. Armato III, A.S. Roy, H. MacMahon, F. Li, K. Doi, S. Sone, and M.B. Altman, "Evaluation of automated lung nodule detection on low-dose computed tomography scans from a lung cancer screening program," Acad. Radiol. 2005; 12:337–346

[200] P. Costello, U.J. Schoepf, K. Peldschus, P. Herzog, S.A. Wood, and J.I. Cheema, "Computer-aided diagnosis as a second reader: Spectrum of findings in CT studies of the chest interpreted as normal reader: Spectrum of findings in CT studies," Chest 2005; 128:1517-1523
[201] Armato et al., The Lung Image Database Consortium (LIDC): ensuring the integrity of expert-defined "truth". Acad-Radiol 2007; 14:1455-1463.
[202] Kobayashi et al., Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs. Radiology 1996; 199:843-848.
[203] A. Jemal, R. Siegel, E. Ward, T. Murray, J. Xu, C. Smigal and M. J. Thun, "Cancer statistics, 2006," CA Cancer J. Clin. 56, 106-130 (2006).
[204] HK Roy et al., Colon Cancer Screening: The Good, the Bad, and the Ugly, 166 Arch Int Med 2177-2179, 2006.
[205] Winawer SJ, Zauber AG, Ho MN, et al. Prevention of colorectal cancer by colonoscopic polypectomy. N. Engl. J. Med. 1993;329:1977-81 (Pickhartd, NEJM).
[206] Pickhardt PJ, Choi JR, Hwang I, et al. Computed tomographic virtual colonoscopy to screen for colorectal neoplasia in asymptomatic adults. N. Engl. J. Med. 2003; 349:2191-2200.
[207] Bond JH. Clinical evidence for the adenoma- carcinoma sequence, and the management of patients with colorectal adenomas. Semin Gastrointest Dis 2000;11:176-84.
[208] Zalis ME, Barish MA, Choi JR, et al. CT colonography reporting and data system: a consensus proposal. Radiology 2005; 236:3-9 (see references 4 and 5).
[209] D Rockey, Colon Cancer Screening, Polyp Size, and CT Colonography: Making Sense of It All?, 131 Gastroenterology 2006-2009, 2006
[210] Kim DH, Pickhardt PJ, Taylor AJ, et al. CT Colonography versus Colonoscopy for the Detection of Advanced Neoplasia. N. Engl. J. Med. 2007; 357:1403-1412.
[211] Smith RA, Cokkinides V, Eyre HJ. American Cancer Society guidelines for the early detection of cancer, 2003. CA Cancer J Clin 2003;53:27-43 (Pickhardt, NEJM).
[212] Summers RM, Yao J, Pickhardt PJ, et al. Computed tomographic virtual colonoscopy computer-aided polyp detection in a screening population. Gastroenterology 2005; 129:1832-1844.
[213] Vining DJ, Gelfand DW, Bechtold RE, et al., "Technical feasibility of colon imaging with helical CT and virtual reality," Am. J. Roentgenol. 1994; 162 (Suppl.):S104.
[214] Pickhardt PJ, Choi JR, Hwang I, et al., "Computer tomography virtual colonoscopy to screen colorectal neoplasia in asymptomatic adults," N. Engl. J. Med. 2003; 349:2191-200.
[215] Kim DH, Pickhardt PJ, Taylor AJ, et al. CT Colonography versus Colonoscopy for the Detection of Advanced Neoplasia. N-Engl-J-Med 2007; 357:1403-1412.
[216] Johnson CD, Dachman AH. CT colonography: the next colon screening examination? Radiology 2000;216:331-41.
[217] Gluecker TM, Johnson CD, Harmsen WS, Offord KP, Harris AM, Wilson LA, Ahlquist DA. Colorectal cancer screening with CT colonography, colonoscopy, and double-contrast barium enema examination: prospective assessment of patient perceptions and preferences. Radiology 2003;227:378–384. (From Summers 2005)
[218] Cotton PB, Durkalski VL, Pineau BC, et al. Computed tomographic colonography (virtual colonoscopy): a multicenter comparison with standard colonoscopy for detection of colorectal neoplasia.[see comment]. JAMA 2004; 291:1713-1719.
[219] Pickhardt PJ, Choi JH. Electronic cleansing and stool tagging in CT colonography: advantages and pitfalls with primary three-dimensional evaluation. AJR Am J Roentgenol 2003;181:799-805.
[220] Zalis ME, Barish MA, Choi JR, et al. CT colonography reporting and data system: a consensus proposal. Radiology 2005; 236:3-9.
[221] A.H. Dachman and M.E. Zalis, "Quality and consistency in CT colonography and research reporting," Radiology 2004; 230:319–323.
[222] Yoshida H, Nappi J, MacEneaney P, Rubin DT, Dachman AH. Computer-aided diagnosis scheme for detection of polyps at CT colonography. Radiographics 2002;22:963–979.
[223] Kiss G, Van Cleynenbreugel J, Thomeer M, Suetens P, Marchal G. Computer-aided diagnosis in virtual colonography via combination of surface normal and sphere fitting methods. Eur Radiol 2002;12:77–81.
[224] Paik DS, Beaulieu CF, Rubin GD, Acar B, Jeffrey RB, Yee J, Dey J, Napel S. Surface normal overlap: a computer-aided detection algorithm, with application to colonic polyps and lung nodules in helical CT. IEEE Trans Med Imaging 2004;23:661–675.

[225] Cathier P, Periaswamy S, Jerebko A, Dundar M, Liang J, Fung G, Stoeckel J, Venkata T, Amara R, Krishnan A, Rao B, Gupta A, Vega E, Laks S, Megibow A, Macari M, Bogoni L. CAD for polyp detection: an invaluable tool to meet the increasing need for colon cancer screening. CARS 2004–Computer Assisted Radiology and Surgery, Proceedings of the 18th International Congress and Exhibition, June 23–26, 2004, Chicago, IL, 2004:978–982.
[226] Kiraly AP, Laks S, Macari M, Geiger B, Bogoni L, Nova CL. A fast method for colon polyp detection in high-resolution CT data screening. CARS 2004–Computer Assisted Radiology and Surgery, Proceedings of the 18th International Congress and Exhibition, June 23–26, 2004, Chicago, IL, 2004:983–988.
[227] Summers RM, Jerebko AK, Franaszek M, Malley JD, Johnson CD. Colonic polyps: complementary role of computer-aided detection in CT colonography. Radiology 2002;225:391–399.
[228] S.A. Taylor, S. Halligan, D. Burling, M.E. Roddie, L. Honeyfield, J. McQuillan, H. Amin, and J. Dehmeshki, "Computer-Assisted Reader Software Versus Expert Reviewers for Polyp Detection on CT Colonography," AJR 2006; 186:696–702.
[229] Shi R, Schraedley-Desmond P, Napel S, et al. CT colonography: influence of 3D viewing and polyp candidate features on interpretation with computer-aided detection. Radiology 2006;239:768–776.
[230] Taylor SA, Halligan S, Slater A, et al. Polyp detection with CT colonography: primary 3D endoluminal analysis versus primary 2D transverse analysis with computer-assisted reader software. Radiology 2006;239:759–767.
[231] Petrick N, Haider M, Summers RM, et al. CT colonography with computer-aided detection as a second reader: observer performance study. Radiology 2008; 246:148-156.
[232] Halligan S, Altman DG, Mallett S, et al. Computed tomographic colonography: assessment of radiologist performance with and without computer-aided detection.[see comment]. Gastroenterology 2006; 131:1690-1699.