SUMMARY MINUTES

CENTER FOR DEVICES AND RADIOLOGICAL HEALTH

RADIOLOGICAL DEVICES ADVISORY COMMITTEE

March 4, 2008

Hilton Washington DC North
Gaithersburg, MD

Radiological Devices Panel:
March 4, 2008
Attendees:

Acting Chairperson:
Leonard M. Glassman, M.D.
Washington Radiology Associates, P.C.

Voting Members:
John D. Bourland, Ph.D.
Wake Forest University

Carl J. D'Orsi, M.D.
Emory University Hospital

Bharat B. Mittal, M.D.
Northwestern Memorial Hospital

Marvin C. Ziskin, M.D.
Temple University School of Medicine

Temporary Voting Members:
Craig K. Abbey, Ph.D.
University of California, Santa Barbara

Donald A. Berry, Ph.D.
The University of Texas, Houston

John A. Carrino, M.D., M.P.H.
Johns Hopkins University School of Medicine

Lori E. Dodd, Ph.D.
National Cancer Institute/NIH

Brian S. Garra, M.D.
Fletcher Allen Healthcare

David Kim, M.D.
University of Wisconsin Medical School

Marilyn Leitch, M.D.
University of Texas Southwestern Medical Center

Otto Lin, M.D.
Virginia Mason Medical Center

Robert D. Rosenberg, M.D.

University of New Mexico Health Sciences Center

Berkman Sahiner, Ph.D.
University of Michigan Medical Center

Kenneth J. Steier, D.O., M.P.H., M.H.A.
Nassau University Medical Center

Daniel R. Swerdlow, M.D.
Georgetown University Hospital

Georgia D. Tourassi, Ph.D.
Duke University Medical Center

A. Christine Watt, M.B., Ch.B.
St. John Hospital

Roy K.H. Wong, M.D.
Walter Reed Army Medical Center

Non-Voting Consumer Representative:
Nancy Finken, M.F. A.
McLean, VA.

Non-Voting Industry Representative:
David K. Spindell, M.D.
Abbott Laboratories

Executive Secretary:
Nancy Wersto

CALL TO ORDER AND PANEL INTRODUCTIONS

Chairman Glassman called the meeting to order at 8:05 a.m. and noted the presence of a quorum. The remaining 2008 meetings are tentatively scheduled for August 12 and November 4. **Executive Secretary Wersto** read the conflict of interest statement. All members and consultants were in compliance, and one COI waiver was issued for Dr. Carrino. **Chairman Glassman** had the Panel members introduce themselves. **Dr. Brogdon** noted that that four Members had recently rotated off the Panel. She thanked them in absensia, and FDA sent them each a letter of commendation and a plaque.

FDA PRESENTATIONS: CAD DEVICES AND MAMMOGRAPHY CAD DEVICES

Dr. Joyce M. Whang, Ph.D., introduced the FDA presentation and an outline of the two-day meeting. The meeting was to cover Computer-Aided Detection and Diagnostic Devices (CAD), with an emphasis on detection devices. The meeting was part of the process of developing a guidance document on the devices. She reminded the Panel that FDA is required to be least burdensome in its expectations of companies while demonstrating safety and effectiveness.

Robert C. Smith, M.D., J.D., presented on the clinical and regulatory background for CAD devices. CAD is intended to aid radiologists with the interpretation of imaging tests. There are four steps of image interpretation: detection, description of the features of the finding, diagnosis, and reporting. CAD devices can be designed to assist with those steps, and the primary purpose is to reduce errors associated with interpretation of imaging results. There are four types of interpretation errors: perceptual (failure to perceive an abnormality), cognitive (mistaking the nature or significance of an abnormality), technical errors, and administrative errors. Technical and administrative errors can cause or contribute to perceptional and cognitive errors.

He then gave an overview of device regulation. He described Class I, class II, and class III devices, the 510(k) approval process for devices with substantial equivalence to a predicate device, and the PMA process. He noted that the Food and Drug Administration Modernization Act of 1997, requires the FDA to use the least burdensome approach. FDA defines "least burdensome" as "a successful means of addressing a premarket issue that involves the most appropriate investment of time, effort, and resources on the part of industry and FDA."

Nicholas Petrick, Ph.D., spoke on general CAD methods. CADe are computer-aided detection devices (also called CAD) which are designed to identify potentially abnormal findings on an image. They are used as prompting devices, and they mark the images to identify potential locations. CADx, computer-aided diagnosis devices, are designed to

process a specific finding and characterize the finding for likelihood of malignancy, to recommend clinical action, or to describe the finding.

CAD encompasses many disciplines, including statistics, pattern recognition, artificial intelligence, physics, image processing, biology, and medicine. The basic blocks of CADe algorithms include acquiring data from digitized film or direct digital devices; image processing by enhancement, segmentation, and features; classification for a single output score; and annotation of potential abnormalities. In a CADx algorithm, there is no need to identify locations. The image processing, feature selection, and classification blocks remain the same.

CAD training is the process of systematically improving performance for a set of data known as the training set. The goal could be maximizing sensitivity or area under the ROC curve. Training can be performed by computer, or humans can tweak or combine parameters. Increased training data increases performance to an asymptotic level and decreases variability. Algorithms are fixed after training.

Clinically, CAD can be used as a first reader, second reader, or concurrent reader. No radiological device has been cleared for first reader use, in which the physician reviews only regions or findings marked by the CAD device. When CAD is used as a second reader device, the physician conducts a complete interpretation without CAD, then reinterprets with the CAD device. With a concurrent read, the physician performs a complete interpretation in the presence of CAD marks. CAD factors influencing clinical use include physical characteristics of the CAD mark, which may cause different responses in different physicians. The introduction of CAD can change interpretations and reading time.

Non-clinical evaluation of CAD algorithms incorporates algorithm description and stability analysis. Algorithm description is important, since different CAD contain different processing. To understand a device, certain information is needed: the target population, the device usage, differences in image processing or segmentation steps, different features or classifiers, and the training and training data used to develop the algorithm. In stability analysis, a stable algorithm is one that produces similar performance with changes in algorithm, features, training, or training databases. Stability increases as the number of training cases increases, the number of initial features decreases, and the complexity of the CAD decreases. Stability analysis also aids in evaluation in future updates of the algorithm.

In clinical testing, CAD algorithms are evaluated on a hierarchical model of efficacy, of which there are six levels, ranging from technical efficacy to societal efficacy. There are two classes of testing: standalone performance testing, which tests the device functionality, and reader performance testing, which tests the performance of physicians using the device and the device's impact on physician performance. In a standalone test, the dataset should represent the target population and target disease condition, but it should be different from the training dataset. A test dataset can be collected by a field test accrual or an enrichment accrual. The Sponsor may want to compare performance of revised algorithms with the same or an expanded test dataset used to train or test the original algorithm. However, this may lead to teaching to the test. The developer may have learned something from the performance of the original CAD on the test data, and the knowledge gained from reusing the dataset may be very small and

unreliable. It may be possible to reuse test data under appropriate constraints to streamline assessment.

Another important component is establishing ground truth. Ground truthing includes whether or not disease is present at the patient level and the location of the disease at the lesion level. Ground truth can be established for cancerous lesions by biopsy and follow-up imaging. For non-cancerous lesions, ground truth can be established by an expert panel. The next component is establishment of scoring rules and methods. One way of scoring is to view the overlap between the CAD and the truth. Another is to measure the distance between the CAD marking and the true centroid. Standalone performance endpoints include lesion-based sensitivity, the number of false positives per image, and the Free Response Receiver Operating Characteristic curve (FROC).

Reader performance testing requires the same steps as standalone testing with the addition of the readers. Readers should be selected to be representative of the intended users. Reader performance testing depends on proper understanding and use of the device as well as proper understanding and implementation of study protocol, so reader training is important. Study endpoints include patient-level analysis, such as sensitivity, specificity, or ROC (Receiver Operating Characteristic) analysis; as well as location-specific analysis, such as location-specific ROC and free-response ROC. Patient level endpoints are used to test CADx devices or CADe when not accounting for location. Location-specific endpoints are used in assessing CADe where location is important or there are multiple prompts on the same image. The truthing rule is critical in location-specific testing. ROC analysis requires correct location of the lesion. Sensitivity can be compared to specificity or diagnoses can be compared with or without CAD. The challenge is to link the measures to clinical relevance and develop a statistical methodology.

In evaluating CAD algorithms, there are prospective and retrospective study designs. A prospective design measures CAD performance as part of actual clinical practice, through field testing. Retrospective reader studies use a dataset of cases collected prior to image interpretation, which are usually enriched. The cases are read offline by one or more readers under specific reading conditions. Dr. Petrick gave examples of the Warren-Burhenne and Multiple Reader Multiple Case (MRMC) study designs. MRMC is more statistically powerful for a given number of cases, while the Warren-Burhenne design can be more difficult to interpret. In MRMC, a set of readers interprets a set of patient images in each of two competing reading conditions, such as with and without CAD. A Warren-Burhenne study consists of two studies: a retrospective study of CAD sensitivity to reduce false negatives and a prospective study of the work-up rate of readers with and without CAD in clinical practice. MRMC testing endpoints can be patient-based or lesion-based.

Thomas E. Gwise, Ph.D., presented on statistical issues in CAD evaluations. Statistical evaluations of diagnostic tests look at both sensitivity and specificity, which are only comparable if they are estimated in the same study. The first question of a diagnostic test is whether or not it provides value: decreased reading time, improved sensitivity and specificity, or improved ROC plot. Most CAD submissions have been labeled as second reader and are to improve the performance of the physician. Because the study conduct

matches the intended use, it is one way to test for a change in performance is to do a multi-center, prospective, randomized clinical trial. Patients would be randomized for CAD assisted and unassisted reading, and follow-up would determine the true disease state. Unfortunately, where disease prevalence is low, such a study would require very large enrollments. Additionally, if patient management depends on the readings, an Individual Device Exemption (IDE) may be required. Retrospective reader studies (commonly with a dataset enriched for stress testing) and MRMC have the benefits of creating no risk for the patients and requiring small sample sizes. However, reader behavior may not reflect routine clinical practice, enrichment can cause biases, and a small number of readers may not generalize.

In a sequential reading study design, the CAD acts as a second reader. Every reader reads every image, first unassisted then CAD-assisted immediately after. Intra-reader variability is minimized, since the assisted and unassisted readings are done close together in time. However, the design is open to reader behavior changes. To minimize the bias, the design can include a washout period between the unaided and the assisted readings. This independent reading session design is more time-consuming and less statistically powerful, but it attempts to control for interpretation bias.

Retrospective reader studies present the problems of enrichment-related biases, choice of controls, reader variability, and how disease localization is addressed. When collecting the study sample, investigators could affect the selection of images to favor one modality over another. This selection bias can be limited by administrative controls and collection of cases from multiple centers. When the sample is enriched with disease positive images, the performance estimates will probably be different than performance in the intended use population, but the differences in performance between modalities may be generalizable to the intended use population. A sample enhanced with too many easily-detected cases may make it difficult to detect a difference in performance between the modalities. A stress test (a study sample enriched with a large portion of positive cases that are difficult to detect) can show the added value of the device, but the performance results cannot be easily generalized across studies. Enrichment can change reader behavior, so investigators should estimate relative performance. A standardized analysis attempts to adjust to a standard population that represents the target population. Direct standardizations weight categories according to their distribution in the standard population. Performance is estimated as an average of category-specific performance. It reduces estimator bias with respect to target population, but it may increase variability.

In CAD studies, the desired effect is the change in performance due to CAD use, so it is often compared to performance of an unassisted single reader. However, studies that require the reader to perform a second read using an image supplemented by CAD do not exclude the possibility of a second look without CAD improving performance. Other studies have compared CAD-assisted reading to a double reading that incorporates a second reader. Reader variability and the reader sample size are important factors. A small number of readers may not be representative of the intended use population. Per patient results cannot be compared to location analyses. The effect size must be considered in the context of study design quality.

Robert C. Smith, M.D., J.D., spoke on mammography CAD devices. The purpose of mammography CAD devices is to reduce errors when interpreting screening

mammograms. The relevant patient and cancer characteristics are cancer size, breast density, finding type, histologic type, and palpability. In screening, 35 percent of cancers are less than or equal to 10 mm, 60 percent less than or equal to 15 mm, and 75 percent less than or equal to 20 mm. The larger cancers are more readily identified and characterized. With regard to breast density, 10 percent of patients have almost entirely fatty breasts, 40 percent have scattered fibroglandular densities, 40 percent are heterogeneously dense, and 10 percent are extremely dense. Greater breast densities are associated with lower sensitivity for detection and higher incidence of interval development following a negative mammogram. Of the types of findings, 30 to 40 percent will be masses, 30 to 40 percent will be microcalcifications, 10 to 20 percent will be a combination of a mass and microcalcification, and 10 to 20 percent will be architectural distortion or focal asymmetry. The histologic type is 70 to 80 percent invasive cancers, 20 to 30 percent ductal carcinoma in situ (DCIS). Regarding palpability, patients undergoing screening will be asymptomatic, but 2 to 5 percent will retrospectively be shown to have symptoms.

There are two types of mammography devices on the market: screen film and digital. DR digital devices use a direct or indirect flat panel. CR digital devices use a photostimulable phosphor. The two standard projections are the craniocaudal (CC) view and the mediolateral (MLO) view. Breast cancer is detected on the basis of four types of mammographic findings: the characteristic morphology of a mass, the shape and spatial configuration of microcalcifications, distortion of breast tissue architecture, and asymmetry between right and left breast. The Mammography Quality Standards Act does not apply to CAD devices. Mammography studies are always interpreted by examination of the CC views from each breast in a side-by-side manner, and likewise for the MLO views. When a finding is identified, the corresponding region on the complimentary view is used to confirm the three-dimensionality of the finding. Comparisons are made to prior mammograms, when available. Mammographic characteristics are reported according to the American College of Radiology (ACR) Breast Imaging Reporting and Data System (BI-RADS) Atlas. A finding is triangulated to identify the three-dimensional location in the breast.

Literature reflects a great variability among mammographers. Sensitivity ranges from 60 percent to 100 percent. Specificity ranges from 35 to 98 percent. The Breast Cancer Surveillance Consortium shows an average sensitivity of 79 percent and average specificity of 90 percent. Mammographic sensitivity is lowest in patients with dense breasts and small masses. Over 99 percent of patients undergoing screening do not have cancer, so practicing radiologists perform analysis very quickly, resulting in possible errors of analysis. A device to reduce errors should focus on the types of cancers radiologists tend to miss. Increased detection should always be weighed against increased false positives.

Over 10 years of mammography screening, 50 percent of women will have at least one false positive. Approximately 20 percent of cancers (half of which are visible) are missed on screening mammography. Of the missed visible cancers, half are missed due to errors of analysis. Most missed visible cancers are masses, architectural distortions, or focal asymmetry; they are small, and they occur in patients with dense breasts. Interpretation errors can be reduced by double reading, which improves

detection but increases the recall rate. CAD devices function as a possible replacement for one of the readers.

In screening, cancer incidence is 4/1000, and CAD devices place at least two marks per patient. For every true positive mark, there will be 249 false positive marks, so it is important to see how this affects the reader.

Four mammography CAD systems have been approved. All devices were first approved for use with digitized versions of screen film mammograms obtained for screening purposes. Supplements have been approved to expand the use to operate on mammograms from full field digital mammography devices (FFDM)and for modified software versions of CAD algorithms. The current indications are "to identify and mark regions of interest on routine screening and [the CC and MLO views of] diagnostic mammograms to bring them to the attention of the radiologist after the initial reading has been completed. … [and to] assist the radiologist in minimizing observational oversights by identifying areas on the original mammogram that may warrant a second review."

The devices were typically approved on the basis of 4 components: standalone performance on missed cancers, standalone performance on cancers detected at screening mammography, standalone performance on normal screening mammograms to determine false positive rate, and screening programs to determine the potential increase in recall rate. Since the first approvals, there has developed a large body of literature on standalone performance and reader performance, employed through retrospective and prospective studies using sequential and historical controls. Standalone testing has shown high sensitivity for calcifications; much lower sensitivity for masses, architectural distortion, or focal asymmetry; and a false positive rate between 2 and 4 marks per patient. Reader performance testing has shown conflicting results for detection of invasive cancers, a trend toward CAD improving a radiologist's detection of calcifications, and an increased recall rate.

Ground truth for cancer is determined by biopsy or surgery. Ground truth for benign findings is determined by biopsy, surgery, or a one-year follow-up mammogram. Ground truth for the location and extent of a finding is determined by an expert panel and can be annotated on the image.

Standalone performance is dependent on case selection, including mammographic characteristics: finding size, pathologic type, breast density, and number of masses versus microcalcifications. Standalone performance testing relies on the method of ground truth determination for the extent and location of disease, the precise scoring metric, and the precise scoring methodology. The larger datasets used in standalone allow for meaningful stratification into relevant subgroups. Measures may be stratified by finding types, pathologic types, size of findings, and breast composition. Overall and stratified standalone performance can be reported per lesion, per view, per breast, or per patient. Without standardized methodologies for case selection, ground truth, scoring metric, scoring methodology, and reporting, it may be difficult or invalid to compare the performance of different devices or to account for differences in detection location and number.

The different characteristics of screen film and FFDM devices may cause different results, both in standalone and reader performance testing. Second reader paradigms increase reading time, and concurrent reading may reduce reading time. Since

CAD has shown sensitivity for calcifications, there may be a clinical role for CAD as a concurrent reader for calcifications.

Chairman Glassman opened the floor for questions. **Dr. Berry** asked if any CADs are set to improve specificity as well as sensitivity. **Dr. Smith** said none had yet been approved. **Dr. Berry** further asked about sensitivity and specificity. **Dr. Smith** said the sequential reader studies have not shown statistically significant improvements. He said the Panel was to give advice on benchmarks for sensitivity and specificity. **Dr. D'Orsi** asked about AUC (area under the curve) data. **Dr. Smith** said none have been submitted. **Dr. Bourland** asked about the DCSC database. Dr. Smith said it is funded by the National Cancer Institute (NCI), uses seven databases linking mammographic and pathologic findings, and that much of the database is publicly available online. All images are obtained from MQSA-certified facilities.

OPEN PUBLIC HEARING: GENERAL CAD

Heang-Ping Chan, Ph.D., spoke for the American Association of Physicists in Medicine (AAPM). AAPM recognizes that CAD will be an indispensable part of diagnostic medicine in the future, that academic research is the driving force of its progress, and that its potential is not yet realized. AAPM urged encouragement and funding of further research. Approval of a CAD system should be conditioned on post-approval evaluations of the system in clinical practice. Standardization of CAD evaluation includes a clear definition of the task, patient population, reader training with CAD, with and without CAD reading design, definition of truth, data analysis methods, identification of biases and variance, and endpoint for assessment of success or failure. Journals should offer fair opportunities for rebuttal or critical review of published studies. Quality assurance procedures should be established for CAD systems implemented in clinical use. Radiologists should get training on interpretation of CAD information before using it clinically. Procedures should be established to ensure that CAD systems are used as labeled. Before a CAD system is used as a first reader, it should be approved for that purpose by the FDA after proper prospective studies.

Stephen Vastagh, MBA, spoke for the Medical Imaging and Technology Alliance (MITA), an organization of manufacturers. MITA maintains and publishes DICOM, a standard on image communication and has worked with FDA on international standardization of medical devices. MITA hoped the meeting would build a common understanding of how CAD products are to be used and evaluated. FDA should evaluate devices based on the sponsors' claims and not require studies on possible off-label use. The definition of CAD does not automatically include improving the radiologists' performance unless the claim is specifically made. MITA recommended that FDA look at the state of the art of CAD and assess the risk of new CAD products in the light of recent development and experience, issue guidance on the separation of Class II and Class III CAD devices (in a collaborative manner that allows industry input), and either expedite the guidance or provide an interim solution.

Robert M. Nishikawa, Ph.D., FAAPM, disclosed financial relationships with FujiFilm, Carestream, and Hologic, but he attended representing himself. Clinical evaluation of CADe is difficult, especially if cancer prevalence is low. Part of the problem is the view that clinical data on CAD screening mammography is conflicting. Dr. Nishawa said that clinical data actually consistently demonstrates the benefit of using CAD. There are 10 studies on the issue. In the longitudinal studies, cancer detection rates increase 1 to 2 percent, 10 percent in the cross-sectional studies. He said that the goal of CAD is not to find more cancers but to find them earlier. For that reason a longitudinal effect is flawed. The goal of CAD is to reduce the false negative rate. Dr. Nishikawa's time expired before he could finish his presentation.

C. Carl Jaffe, M.D. of the National Cancer Institute spoke as an individual and had no financial interests to report. He said expert observers are good but inconsistent and effective technology can increase performance. He pointed to the variation in sensitivity among radiologists and said that readers who would volunteer for the study may be a self-selecting, biased sample set. On quantitative monitoring of change over time, observer performance is mediocre to poor. Ground truth is often unknowable. One benchmark is comparative human performance. Validation is needed, so large databases are needed. NCI encourages the development of effective CAD by developing open, public databases with metadata and clinical data. The unresolved questions are how large the reference image set should be, what the training and testing components are, who possesses the database, and who judges performance and by what metric.

Steve Worrell from Riverain Medical said his company was the only one to demonstrate safety and effectiveness of a CADe device intended to identify and mark regions of interest on frontal chest radiographs. The device is intended as a second reader. Riverain strongly believes that Class III regulation is necessary to mitigate risks associated with chest CAD devices. CAD devices are currently regulated as both Class II and Class III devices, and the distinction is unclear. FDA must clearly delineate between Class II and Class III devices. Riverain advocates requiring a reader study consistent with intended use for initial chest CAD PMA applications. Further modifications could be supported by standalone testing, as is consistent with the Least Burdensome provisions.

Akira Hasegawa, Ph.D., of FujiFilm spoke on risk assessment of CAD. He said there are many types of CAD with different indications for use (IFU) and different risks. These risks can be assessed by comparing them with the physicians' standard reading procedures without CAD. A second reading is used to correct for human error, and there are CAD detection devices to support that process, Type 1 devices. The second reading does not affect the standard reading process and provides no diagnostic information. Type 2 CAD devices assist the radiologist in interpretation and are diagnosis devices. These devices carry risk because they affect the user's decision-making. A Type 3 device is a CAD for concurrent read. It affects the user's initial searching process but does not provide diagnostic information. Its risks include a negative effect on users' searching and possibly causing the users' satisfaction of search. Dr. Hasegawa's time expired before he finished his presentation.

Dr. D'Orsi asked about use of Types 1 and 2 on screening exams. **Dr. Hasegawa** said that would depend upon the intended use. **Dr. D'Orsi** asked if the database for the ACRIN study would be available to the FDA. **Dr. Jaffe** said that ACRIN is a grantee and arrangements for data-sharing are being worked on. There are multiple parties with financial interest in the studies, so the issue of access is complex. He said open databases would help the field.

OPEN PUBLIC HEARING: MAMMOGRAPHY

Rachel Brem, M.D., of iCAD Medical said there is a well-established, rigorous clinical testing paradigm for approval of mammography CAD. Data presented with iCAD's PMA submission showed a 20 percent improvement in breast cancer detection and detection 15 months earlier with CAD. The increase in recall rate was 0.5 percent, a statistically insignificant amount. The data utilized a large number of breast cancers to establish the stability of the data and the false positive rate. For film screen and digital mammograms, only confirmatory studies are needed to show comparable performance based on the original PMA submission. The increase in recall rate is necessary to identify more cancers. Mammography CAD is approved using a well-established testing paradigm. Scholarly literature on CAD demonstrates positive impact on radiologist performance in a variety of conditions. With the large societal burden of breast cancer, the improvement in cancer detection with CAD is clinically significant. The current clinical testing paradigm is appropriate to ensure the safety and effectiveness of mammography CAD.

Julian Marshall of Hologic said mammography CAD is clinically useful. There are many, large, and varied studies on CAD, but there are questions as to what will establish the clinical efficacy. In the last few years, prospective clinical studies have demonstrated an increased cancer detection rate. Mammography CAD was approved three years before the first prospective independent study was published. Requiring such studies before approval will delay getting the devices to clinicians. Prospective studies show an increase in detection with CAD use because cancers are discovered earlier. He urged that upgrades to approved CAD devices be tested using assessments of standalone performance.

Robyn L. Birdwell, M.D., was sponsored by the Society of Breast Imaging. She said CAD reduces false negatives with a modest increase in the recall rate. CAD is necessary because radiologists can miss cancers during screening, and the benefit is equivalent, using film or digital methods. The increase in recall rate is necessary in order to treat more cancers. The increase in detected cancers using CAD is 9.7 percent. Depending on the reader, other studies have shown higher increases. The peer-reviewed literature of prospective, sequential read clinical studies and historical controlled clinical studies show that FDA-approved CAD systems and algorithm improvements are efficacious and safe.

Gillian M. Newstead, M.D., FACR, of the American College of Radiology disclosed financial relationships with Philips Medical, CME, and Hologic. She said that with the increase in information available and considered important, radiologists are overwhelmed

and need computer assistance to view, manipulate, and analyze large, complex imaging datasets. With breast MR images, radiologists are aided by having computers depict the kinetic characteristic lesions rather that visually estimating them. Computers will play an increasingly important role in the diagnostic assessment of lesions as technology continues to give larger datasets.

Dr. Nishikawa returned to speak on clinical studies of CADe for screening mammography. He noted the 9.7 percent increase in sensitivity and the 13 percent increase in recall rate. He said the goal is to reduce the number of missed cancers, and it is reasonable to expect CADe to work like a double reading with 2 radiologists, which produces a 10 percent increase in sensitivity and a an increased recall rate comparable to CAD. He said the inclusion of benign cases in evaluations is an added complication, since they would look like cancer to a radiologist as well. He said that requiring an observer study for the same CADe running on images from different digital systems is unreasonable. He said that it is more reasonable to approve CADe on one FFDM system, then require equivalence or better from other systems.

Maryellen L. Giger, Ph.D., FAAPM, of the University of Chicago disclosed financial relationships with NIH, the US Army, and R2/Hologic. She said that the progress of CAD depends on research funding, careful clinical studies, and training with the methods when introducing them into the clinical arena. Many factors affect radiologists' performance levels, and interpretation can be done by the radiologist, double reading, or radiologist and CAD. Double reading improves detection sensitivity with some increase in the recall rate. The sensitivity increase and the recall rate increase is comparable to the increases from single read with CAD. It might be useful to demonstrate equivalency between single read with CAD and double read. The way to do this would be to determine a performance standard for CAD based on published data, require a CAD to meet that standard, perform the cooperative study, and only allow radiologists trained in CAD usage to participate. After equivalency is demonstrated, all changes or improvements to the system must meet or exceed the earlier specified standard. An independent technology assessment institute with a sufficiently large database with appropriate distributions of cancer types could be tasked with performance assessment.

Dr. D'Orsi asked about digitized screen film images. **Dr. Nishikawa** said that they are probably equivalent to digital data. **Dr. Brem** added that efficacy and safety were established on digitized analog data and that studies show comparable performance is needed for different image sources. **Dr. Bourland** asked about the independent center Dr. Giger suggested. **Dr. Giger** said it would be an independent center that preserved the integrity of the database. **Dr. Dodd** asked about a reasonable FDA endpoint threshold. **Dr. Nishikawa** said it would depend on the underlying shape of the ROC curve and that more investigation is needed. .

PANEL DISCUSSION: MAMMOGRAPHY CAD DEVICES

Chairman Glassman commented that the distinction between CADe and CADx is artificial, since detection and diagnosis exist on a continuum. **Dr. Garra** said that a

CADx device would, theoretically, diagnose cancer as well as other lesions. **Dr. Tourassi** made the distinction between CADe and CADx as the decision to be made. **Dr. D'Orsi** said the separation between CADe and CADx is critical, since a mixture of the two could cause reader bias. He added that FDA should focus on how many cases go into the CADx training algorithm. **Dr. Mittal** noted that the test data includes patients with architectural distortion due to radiation plus surgical resection. **Dr. Wong** expressed interest in developing a regulatory process in which CAD could be tested in an unbiased manner. Larger databases will lead to more sophisticated testing. **Dr. Abbey** said that reader performance results get to the utility of the decision. **Dr. Garra** asked if FDA intended to make more stringent criteria for approval. He commented that he supported funding to help NIH build a generalized, uniform, carefully laundered database of high quality images. **Ms. Brogdon** commented that FDA had no plans to make criteria stricter but was looking for recommendations. **Dr. Watt** stressed the importance that practicing radiologists not mistake CADe for CADx. **Dr. Berry** said it is important that companies show improved sensitivity. She added that enrichment can compensate for sample size and commented on the inapplicability of recall rate as an endpoint. **Dr. Tourasi** said that there must be standardization and an emphasis on training the end users. **Ms. Finken** expressed concern for the anxiety caused by a call-back and urged an approach to call-backs that will not scare the patient away. **Dr. Spindell** asked about the risks of a false positive as opposed to the risk of a false negative and which a patient would prefer. He further asked if follow-up CAD was considered equivalent to a double reading. He said a universal dataset would level the playing field. **Dr. Kim** said that training and labeling are important, since the devices must not be used for primary reading. He further commented that the test database must be large and standardized. **Dr. Leitch** said that blending detection and diagnosis remains an issue. It is important to consider whether CAD will reduce or increase the cost of screening. The increased sensitivity and ability to detect ductal carcinoma are important. She commented that the newer digital images show calcifications better, so she wondered if CAD added as much to the digital images. **Dr. Sahiner** said that data submitted for approval should indicate the expected clinical effect. That may be impractical due to resource limitations, but it is important to look at the clinical effect. **Dr. Carrino** said the unit of analysis should probably be per lesion and that the testing should be done with a reader, since that is the intended use. **Dr. Rosenberg** commented that large databases would be helpful but are difficult to obtain. **Dr. Dodd** urged steering away from prevalence measures such as recall rate when comparing across databases or studies. She agreed that sensitivity and specificity are appropriate, but reader variability should be considered. She said ROC analysis is appropriate, though it is difficult to determine what AUC improvement represents a clinically significant improvement. She noted that CAD might have less impact on more experienced readers. She expressed concerns of the validity of comparisons across databases. **Dr. D'Orsi** said that mammographic, colon, and chest CAD cannot be generalized. The lesions are generally small, as is the signal to noise ratio. More data is needed on what CAD misses. He said the standardized data set should be retrospective, since that allows for better control of biases. **Dr. Lin** expressed concern about the temptation to rely on CAD. He said standalone testing is a surrogate endpoint. **Dr. Bourland** said the main issue is the standardized databases. He noted the existence of quality assurance issues and the importance of training. The standardized database

Deleted: Sahinger

should match the indications stated by the manufacturer. **Dr. Steier** wondered about quality control and variability among products.

Chairman Glassman noted that the Panel felt CAD is important but that it may be difficult to assess CAD's role. There was a general consensus that a standardized database for testing should be developed, that there was a continuum between CADe and CADx, that the sensitivity to specificity in ROC were not satisfactory. He commented that the way to evaluate effectiveness might be to test the device with the least skilled readers.

M1. Please discuss the role of standalone performance testing in the clinical evaluation of mammography CAD devices.

a. If you believe standalone testing should be requested in the evaluation of these devices, please, provide your recommendations or comments on:

i. The merits of per lesion, per view, per breast and per patient endpoints;

ii. Whether certain substrata, for example, mammographic finding type, size, breast composition or others should be considered in device testing and labeling; and

iii. What marking or scoring methodology should be used for reporting findings?

b. If you believe that there are specific situations where standalone performance testing may not be important, please, comment on what those might be.

Panel consensus was that standalone testing is necessary but not sufficient for the evaluation of a product and that it gives an unbiased look at the function of the product without the confounding effects of the end user, which is valuable. The Panel felt that per lesion, per view, per breast and per patient endpoints were useful but that per lesion and per view were most useful. The data for the substrata is important for analysis and labeling. The Panel preferred a scoring using the distance of the centroids, and the marks should be large and stand out. The Panel could not think of an instance in which standalone performance testing would not be important. While changes in the algorithm should require testing, they did not decide exactly where the line between standalone only and more extensive testing should be drawn. Dr. Berry added that standalone testing should not be the primary endpoint of a pivotal study.

M2. Please discuss the role of reader performance testing in the clinical evaluation of mammography CAD devices.

a. If you believe reader performance testing should be considered in the evaluation of these devices, please provide your comments or recommendations on:

i The appropriate primary endpoint(s) and corresponding clinically significant effect sizes, specifically to comment on the use of ROC analyses;

ii The merits of per lesion, per view, per breast and/or per patient endpoints in the assessment of the endpoints;

iii Whether effectiveness analyses should be conducted separately or not for cancers manifesting as masses versus microcalcifications;
iv Whether reading time should be assessed, and if so, how.

b. If you believe that there are any specific situations where reader performance testing may not be necessary, please, comment on what those might be.

Panel consensus was that reader studies should be the primary analysis, that per patient endpoints are important but that per lesion and per view endpoints should not be ignored. The Panel said ROC analyses were useful, but area under the curve may not be meaningful in itself. Panel consensus was that effectiveness should be conducted for cancers with different findings. The Committee felt that reading time is an important factor, at least for labeling. Standalone testing would probably be sufficient for minor modifications.

M3. Please discuss whether there are other types of performance testing you believe should be considered in the clinical evaluation of mammography CAD devices.

Panel consensus was that retrospective testing, both in reader and standalone, was sufficient.

M4. The prevalence of breast cancer cases in a screening population is relatively low. Please, provide comments on the practice of using an enriched dataset for the clinical evaluation testing discussed in M1-M3.

a. If you believe that an enriched dataset may be used for these evaluations, please, discuss what you believe to be the appropriate clinical and mammographic characteristics (or range of characteristics). Please, consider whether the following characteristics of the screening population should be considered when designing an enriched database or stress test:

i. Breast density: 40-50% of patients with heterogeneously dense or extremely dense breasts;
ii. Proportion and types of masses and microcalcifications, approximately, evenly distributed with a sufficient number of additional patients with architectural distortion alone;
iii. Size and palpability for cancers: non-palpable and a majority with size < 1.0 cm.
iv. Distributions of microcalcifications: small clusters of up to five microcalcifications for a third of the cases, and;
v. 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, et cetera.

Deleted: etcetera

In addition, please comment on whether the expected effect size should be adjusted if an enriched dataset is used. If so, how and why?

b. If you believe that enrichment is inappropriate, please provide your reasons and whether there would be an alternative method of assessing these devices in light of the low prevalence of disease.

Panel consensus was that the Committee recommends an enriched dataset. For standalone testing, the enrichment can be stressed and can be a large set, since there is no concern about reader bias. For the reader sets, however, the enrichment should mirror the normal distribution of cancers in the imaging population with, if possible, prior negative mammograms from patients who developed cancer that was visible a year later. The Panel said that all of the above categories should be included and that there is no alternative to enrichment.

M5. Mammograms obtained on FFDM devices have characteristics that are strongly dependent on engineering design and device hardware and software. If a mammography CAD has been approved to operate with screen-film or a specific FFDM device(s), what data should be used to assess its performance with the different FFDM device? Would one or the other suffice? Are there other types of studies that should be provided instead of or additionally?

The sense of the Committee was that they did not know enough about image processing and the link between CAD and FFDM. If CAD is late in the process, standalone testing might be sufficient. However, early in the digital process, the Committee could not be sure it runs like film, and reader studies may be needed. Dr. Sahiner noted that images from different FFDM manufacturers may be more similar earlier in the process, since it is the processing that is unknown. .

M6. FDA does not specify indications for use, but reviews indications for use that are requested by companies. What are the Panel's views regarding second reader versus concurrent reading using a CAD device? Specifically,

a. How are mammography CAD devices used clinically?

b. Are second reader and concurrent reading modes both clinically relevant options for use in practice? If not, which paradigms are appropriate for mammography CAD devices?

c. Do you believe users understand that if a device is labeled as a second reader, they (i.e. the physician) should always read the radiological image completely before turning to the CAD?

Panel consensus was that clinical use was both concurrent and sequential and that secondary reading was preferable, based on the data and the labeling. Under certain circumstances, concurrent may be acceptable, but it is not the best practice, and training is the best way to promote the best practice. Ms. Brogdon asked if the labeling should conform to the real use of the devices or to FDA's approvals. Dr. D'Orsi said the data indicates against concurrent use and suggested stronger language. Ms. Brogdon asked how a proposed concurrent reader should be studied. The Panel said a reader study based on that paradigm, comparing concurrent reading to reading with no CAD. .

M7. FDA has provided you with a bibliography of the published literature for mammography CAD devices. Please discuss whether these publications provide us with any additional information as to how such devices should be evaluated in the future.

Panel consensus was that the papers offered postmarket information but nothing on more effective ways to evaluate the devices pre-market. Dr. Garra said the studies showed problems generated by trying to get statistically significant changes.

FDA PRESENTATION: COLON CAD DEVICES

Frank Samuelson presented on colon CAD devices. Colon cancer is believed to arise primarily from adenomatous polyps. Most polyps are hyperplastic, not adenomatous. Larger polyps are more likely to be cancerous. About 10 percent of adenomatous polyps larger than 1.5 centimeters may contain invasive cancer. About 8 percent of the screening population has a polyp larger than 1 centimeter. Polyps are relatively easy to remove. Colorectal screening is recommended for people over 50 years of age.

Diagnostic methods include barium enemas; fecal blood testing; optical colonoscopy (OC), the current standard; and CT colonography (CTC). Both OC and CTC include bowel cleansing and insufflation. CTC uses two CT scans, supine and prone. CTC is less invasive and requires no sedation, but it does expose the patient to an X-ray dose. OC provides the ability to perform a resection or polypectomy.

In CTC, the supine and prone scans are rendered as slices and in 3D. The clinician's report will include polyp sizes, locations, morphology, and a summary diagnostic assessment. One method of reporting is the C-RADS system, which includes a rating scale with 0 as an inadequate study and 1-4 as increasing likelihood of malignancy. Detection rates are dependent on polyp size. Sensitivity increases with size, while specificity remains constant. CAD devices mark suspected polyp locations on a CT work station display, and the doctor determines the significance. The potential benefits are improved polyp detection, reduced reading times, and guidance for OC.

In implementation, the organs are segmented and the surface of the colon digitally created, the vertex features are calculated, the software calculates sphericity and curvature, vertices are grouped by feature to create regions of interest (ROI), and the CAD algorithm decides what ROI should be marked. .

Most CAD devices mark multiple locations that are not polyps. FROC curves are often used to demonstrate performance. CTC CAD devices generally have sensitivities of around 90 percent for polyps greater than 10 mm. For stand-alone performance studies, there are questions of whether a mark on the supine scan but not the prone constitutes a true positive or what overlap criteria should be required on CAD marks to count as a true positive.

In retrospective CTC CAD reader studies, there are questions of how the radiologist's performance is affected by CAD. The studies tend to be MRMC with enriched datasets. All studies showed increased sensitivity and decreased specificity. The primary reading can be in 2D or 3D, and CAD devices may be more effective in one or the other. The concurrent reader may save time over second reader, but it is still more

time-consuming than unaided reading. Second reader may be more sensitive than concurrent reading. There are no known prospective clinical studies on the effects of CTC CAD.

CAD output is dependent on patient parameters, CT parameters, and image processing. OC has been used as a reference standard for ground truth, but it misses 11 percent of lesions, some of which CTC finds. OC and a CTC report has been used, but lesions can still be missed. OC can be combined with an expert review, but there is variability among experts. Study endpoints can be sensitivity and specificity or a summary statistic. The unit of measure can be by the patient or by the polyp, and there is the question of which polyps are relevant: all polyps, adenomatous polyps, or polyps of a certain size.

OPEN PUBLIC HEARING

Ronald M. Summers, M.D., Ph.D., an NIH senior investigator, disclosed relationships with iCAD and Viatronix. He raised four issues: patient selection, image acquisition, performance benchmarks, and reading paradigm. He said that most sensitivities of CTC CAD systems are not accurate for estimating performance in the clinic due to selection bias, inadequate use of common databases, and training and testing on the same data. He said that characteristics of the dataset must be clearly defined with detailed specifications. There must also be a separate test set and an external validation. He said the labeling should be specific, specifying scanning parameters and the preferred bowel preparation. The proven protocols should be used. The imaging parameters should focus on polyps 6mm and larger. Performance benchmarks should be laddered by the size of the polyp. The median false positive rate should be 8 per patient, and report sensitivities should be based on all polyps retrospectively visible by CTC. He recommended the second reading paradigm, said that first read should be discouraged by requiring the clinician to record the pre-CAD read. He concluded that standardization and high benchmarks will lead to effective CAD systems.

Maha Sallam, Ph.D. of iCAD said that clinical testing should be based on indications for use and that standalone as well as potential impact on readers is needed. CAD systems are typically used in a second read scenario, though other paradigms have been suggested. The second read allows for adding detections without impacting the initial full review, but testing is needed to measure incremental impact on performance. Concurrent read results in a more complex interaction between the reader and CAD findings, and testing is needed to compare reading with and without CAD. A first read scenario requires rigorous stand-alone performance testing and a comparison between reader performance with and without CAD. It is important that the sponsors indicate the proposed reading paradigm.

Gareth Beddoe, M.D., of Medicsight PLC read a statement from his company, which manufactures colon CAD devices. Medicsight welcomes clarification from the FDA on classification and guidance on appropriate studies to support clearance. CADe and CADx present different risks. Colon CAD indications require that the radiologist review all images, not just those identified by the software. Medicsight requested that FDA

classify different types of CAD differently and that FDA limit the requirements for data to those that satisfy the labeled intended use. He read a discussion of the guidelines for statistical analysis for MRMC studies. Medicsight requested that FDA consider alternative approaches to MRMC ROC analysis for studies of CTC CAD, since the ROC analysis is inappropriate and questionable. The vast majority of polyps are benign. Confidence scores regarding the likelihood of malignancy cannot be applied meaningfully to colon polyps. Because the confidence scores will be skewed, the ROC will not work. Colonoscopy without prior imaging is the most commonly used method for colorectal cancer screening. False positives will lead to OC, but without CTC, the patient would have OC anyway. ROC analysis is inappropriate for CTC studies for conceptual, statistical, practical, and ethical reasons. The company asked the Panel and FDA to recognize that alternative analysis plans for MRMC studies are appropriate and asked that FDA consider alternatives to ROC analysis.

Chairman Glassman asked **Dr. Hasegawa** to return and complete his presentation. He resumed with his discussion of Type 3 CAD, which is for a concurrent read, and showed a decision tree for assessing risk. The first question was, "human makes final diagnosis/decision?" Any device that makes the decision itself is not CAD but an automated detection or diagnosis device. If the images are not reviewed by a human, it is a computerized screening device, not CAD. He summarized that different CAD have different IFU and different risk factors. Since the three different types of CAD are completely different devices, they should be discussed separately.

Dr. Berry asked Dr. Samuelson to talk about sensitivity and specificity. **Dr. Samuelson** said that a question before the Panel is to what CAD should be sensitive, polyps or cancer. It is extremely difficult for a radiologist or colonoscopist to tell a malignant from a benign polyp by visual inspection. **Dr. Steier** asked about requiring a pre-read by the clinician before he or she could progress to the CAD component. **Dr. Summers** said slippage into the first reader paradigm could be prevented if radiologists were required to record the pre-CAD reading, and those records could be audited for accreditation. Dr. **D'Orsi** asked about ROC's irrelevance. **Dr. Beddoe** said he did not prepare a statistical argument and could not comment on it.

PANEL DISCUSSION: COLON CAD DEVICES

Dr. Wong said colon CAD is a major advance and that at Walter Reed the patient can choose between CTC and OC. The hope is that CAD can become a primary reader, in order to save time. **Dr. Swerdlow** agreed, noting the length of time a CTC reading takes. Finding one large enough polyp will trigger an OC, which would catch all polyps, but the first reader must be trustworthy. **Dr. Kim** said CAD will have an interesting and positive role in CTC. The sensitivity of the technique lies in redundancy, and CAD will add a layer of redundancy. The set point for CAD should be large to catch the important lesions. Measuring per polyp, CTC will have a lower sensitivity for a given size, compared to per adenoma. Colorectal cancer screening looks for polyps that usually reach a certain size before cancer occurs, so size of one centimeter is a logical surrogate endpoint. **Dr. Lin** said ROC curve analysis is unneeded, since the colon is different from

the breast. The decision is binary: whether or not the patient needs OC. He suggested 6 mm as the cutoff point. With a set cutoff point, sensitivity and specificity can be calculated without ROC curve analysis. Ground truth can be established with combined information from OC and the virtual colonoscopy. For CTC, the prevalence of lesions of interest is higher than for breast cancer, so enrichment may not be needed. CAD should be a second reader, since there are still questions as to CTC accuracy. The overlap criterion is more difficult, due to the mobility of the colon. **Dr. D'Orsi** asked if there were enough staff to handle screening at the level of eligible people. **Dr. Kim** said there is sufficient hardware but insufficient staff. CAD will be valuable if radiologists are in a hurry. **Dr. Dodd** asked about the ability to assign the likelihood of a polyp being greater than a certain size. **Dr. Kim** said Walter Reed used the C-RADS classification, which scored for the likelihood of a perceived polyp being a true soft tissue polyp. **Dr. Berry** noted that with colon screening there is not the anxiety caused by false positives. **Dr. Garra** said that ROC analysis is appropriate in the detection of colon cancer. **Dr. Tourassi** asked about reader variability with CTC. **Dr. Kim** said studies have variously shown sensitivity to detect a 10 mm polyp from 93 percent to 50-60 percent. He said the largest source of variation was between 2D and 3D; other issues were procedures and training. The consensus is 90 percent sensitivity.

C1. Please discuss the potential clinical utility of CTC colon CAD. Possibilities to consider include: improved sensitivity to detection of polyps of different sizes; reduced reading times; and guiding optical colonoscopy for intervention.

Panel consensus was that there is evidence of improved sensitivity to detect polyps using CAD over virtual colonoscopy without CAD, as well as evidence of decreased specificity. The critical size is between 6 and 10 mm. CAD should be a second read, so there will be no reduced reading time. The virtual examination, not CAD, guides colonoscopy.

C2. Establishing ground truth (i.e, whether disease is present, and if so, its location and extent) is crucial for the evaluation of the performance of any CAD device. Please provide your recommendations for defining ground truth for colon CAD devices.

Panel consensus was that for evaluation of CAD, that an optical colonoscopy for positive cases and follow-up virtual colonoscopy for discordant cases and an expert panel for negative cases is adequate for determining ground truth. Dr. Garra said that there could be anchor points of patients who have had the procedure twice.

C3. Please discuss the role of standalone performance testing in clinical evaluation of colon CAD devices.

> **a. If you believe standalone testing should be requested in the evaluation of these devices, please, provide your recommendations and comments on whether certain substrata (e.g., polyp size, shape, pathology or location; comorbidities; CT dose or imaging protocol; or others) should be considered in device testing and labeling.**

b. If you believe that there are specific situations where standalone performance testing may not be important, please, comment on what those might be.

Panel consensus was that standalone testing is important and that polyp size, 6 mm and larger, should be considered. CT dose and imaging protocol must be known to ascertain whether they are clinically relevant. The dataset should be enhanced with polyps of varying locations, including the flexures, which may be more difficult to find; and flat polyps, which are also difficult to find. The demographics of the test set should be clinically appropriate with the usual patient population. Standalone testing is important in all instances. The comorbidity of diverticulosis should be considered.

C4. Please, discuss the role of reader performance testing in the clinical evaluation of colon CAD devices.

a. If you believe reader performance testing should be considered in the evaluation of these devices, please provide your comments and recommendations on:

i. The appropriate primary endpoint(s) and corresponding clinically significant effect size(s)? Please specifically comment on the use of ROC analyses;

ii. The merits of per lesion, per segment and/or per patient endpoints in the assessment of endpoints;

iii. Whether reading time should be assessed, and if it, how?

b. If you believe that there are specific situations where reader performance testing may not be necessary, please, comment on what those might be.

Panel consensus was that reader performance testing should be done. Clinically effective sizes are 6 millimeters and greater. ROC analysis is appropriate, as is FROC analysis. General consensus was that the endpoint should be per lesion, rather than per segment or per patient, but if ROC analysis is desired, it would be converted to a per segment analysis. Reading time should be assessed. When there are algorithm changes or performance testing, reader performance testing may not be necessary. Standalone testing would be adequate, but for major changes the requirements would be similar to mammography CAD. Chairman Glassman said that this is the least burdensome approach.

ADJOURNMENT

Due to the time, **Chairman Glassman** said that questions C5-C7 would be addressed on the following day. The meeting adjourned at 5:58 p.m.

I certify that I attended this meeting on March 4, 2008 and that these minutes accurately reflect what transpired.

_____

Nancy Wersto
Executive Secretary

I approve the minutes of this meeting as recorded in this summary.

_____

Leonard M. Glassman, M.D.
Acting Chairman

Summary Prepared by

Eric M. Hendrixson
Neal R. Gross & Company
1323 Rhode Island Ave, NW
Washington, DC 20005
(202) 234-4433
Wednesday, March 4, 2008