Guidance for Industry and FDA Staff

Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Approval (PMA) and Premarket Notification [510(k)] Submissions

Document issued on: July 3, 2012

The draft of this document was issued on October 21, 2009.

For questions regarding this guidance document, contact Nicholas Petrick (OSEL) at 301-796-2563, or by e-mail at Nicholas.Petrick@fda.hhs.gov; or Mary Pastel (OIVD) at 301-796-6887 or by e-mail at Mary.Pastel@fda.hhs.gov.

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Devices and Radiological Health

Division of Imaging and Applied Mathematics
Office of Science and Engineering Laboratories

Division of Radiological Devices
Office of In Vitro Diagnostic Device Evaluation and Safety

Preface

Public Comment

You may submit written comments and suggestions at any time for Agency consideration to the Division of Dockets Management, Food and Drug Administration, 5630 Fishers Lane, rm. 1061, (HFA-305), Rockville, MD 20852. Identify all comments with the docket number listed in the notice of availability that publishes in the *Federal Register*. Comments may not be acted upon by the Agency until the document is next revised or updated.

Additional Copies

Additional copies are available from the Internet. You may also send an e-mail request to dsmica@fda.hhs.gov to receive an electronic copy of the guidance or send a fax request to 301-847-8149 to receive a hard copy. Please use the document number (1698) to identify the guidance you are requesting.

Table of Contents

Contains Nonbinding Recommendations

Guidance for Industry and FDA Staff

Clinical Performance Assessment: Considerations for Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Approval (PMA) and Premarket Notification [510(k)] Submissions

This guidance represents the Food and Drug Administration's (FDA's) current thinking on this topic. It does not create or confer any rights for or on any person and does not operate to bind FDA or the public. You can use an alternative approach if the approach satisfies the requirements of the applicable statutes and regulations. If you want to discuss an alternative approach, contact the FDA staff responsible for implementing this guidance. If you cannot identify the appropriate FDA staff, call the appropriate number listed on the title page of this guidance.

1. Introduction

This guidance document provides recommendations to industry, systems and service providers, consultants, FDA staff, and others regarding clinical performance assessment of computer-assisted detection (CADe[1]) devices applied to radiology images and radiology device data (often referred to as "radiological data" in this document). CADe devices are computerized systems that incorporate pattern recognition and data analysis capabilities (i.e., combine values, measurements, or features extracted from the patient radiological data) intended to identify, mark, highlight, or in any other manner direct attention to portions of an image, or aspects of radiology device data, that may reveal abnormalities during interpretation of patient radiology images or patient radiology device data by the intended user (i.e., a physician or other health care professional), referred to as the "clinician" in this document. We have considered the recommendations on documentation and performance testing for CADe devices made during the public meetings of the Radiology Devices Panel on March 4-5, 2008[2] and November 17-18, 2009.[3] We have also considered the public comments received on the draft guidance announced in the *Federal Register* on October

[1] The use of the acronym CADe for computer-assisted detection may not be a generally recognized acronym in the community at large. It is used here to identify the specific type of devices discussed in this document.

[2] https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfAdvisory/details.cfm?mtg=694

[3] http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/MedicalDevices/MedicalDeviceAdvisoryCommittee/RadiologicalDevicesPanel/UCM197419.pdf

21, 2009 (74 FR 54053).

FDA's guidance documents, including this guidance, do not establish legally enforceable responsibilities. Instead, guidance documents describe the Agency's current thinking on a topic and should be viewed only as recommendations, unless specific regulatory or statutory requirements are cited. The use of the word *should* in Agency guidance documents means that something is suggested or recommended, but not required.

2. Scope

This document provides guidance regarding clinical performance assessment studies for CADe devices applied to radiology images and radiology device data. Radiological data include, for example, images that are produced during patient examination with ultrasound, radiography, magnetic resonance imaging (MRI), computed tomography (CT), positron emission tomography (PET), and digitized film images. As stated above, CADe devices are computerized systems intended to identify, mark, highlight, or in any other manner direct attention to portions of an image, or aspects of radiology device data, that may reveal abnormalities during interpretation of patient radiology images or patient radiology device data by the clinician.

By design, a CADe device can be a unique detection scheme specific to only one type of potential abnormality, or a combination or bundle of multiple parallel detection schemes, each one specifically designed to detect one type of potential abnormality revealed in the patient radiological data. Examples of CADe devices that fall within the scope of this guidance include:

- a CADe algorithm designed to identify and prompt microcalcification clusters and masses on digital mammograms,
- a CADe device designed to identify and prompt colonic polyps on CT colonography studies,
- a CADe designed to identify and prompt filling defects on thoracic CT examination and,
- a CADe designed to identify and prompt brain lesions on head MRI studies.

We may consider a CADe algorithm to be modified, for the purposes of this guidance, if, for example, a change has been made to the image processing components, features, classification algorithms, training methods, training data sets, or algorithm parameters. See the guidance entitled **Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions**[4] for information on CADe 510(k) submissions, including when a clinical performance assessment may be recommended for a CADe 510(k) submission. Contact the Agency directly for advice on CADe PMA submissions.

This guidance does not cover clinical performance assessment studies for CADe devices that are intended for use during intra-operative procedures or for computer-assisted diagnostic devices (CADx) and computer-triage devices, whether marketed as unique devices or bundled with a CADe device that, by itself, may be subject to this guidance. Below is further explanation of the CADx and computer-triage devices not covered by this guidance:

---

[4] http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm187249.htm

- CADx devices are computerized systems intended to provide information beyond identifying, marking, highlighting, or in any other manner directing attention to portions of an image, or aspects of radiology data, that may reveal abnormalities during interpretation of patient radiology images or patient radiology device data by the clinician. CADx devices include those devices intended to provide an assessment of disease or other conditions in terms of the likelihood of the presence or absence of disease, or devices intended to specify disease type (i.e., specific diagnosis or differential diagnosis), severity, stage, or intervention recommended. An example of such a device would be a computer algorithm designed both to identify and prompt potential microcalcification clusters and masses on digital mammograms, and to provide a probability of malignancy score to the clinician for each potential lesion as additional information.
- Computer-triage devices are computerized systems intended to in any way reduce or eliminate any aspect of clinical care currently provided by a clinician, such as a device for which the output indicates that a subset of patients (i.e., one or more patients in the target population) are normal and therefore do not require interpretation of their radiological data by a clinician. An example of this device is a prescreening computer scheme that identifies patients with normal MRI scans that do not require any review or diagnostic interpretation by a clinician.

For any of these types of devices, we recommend that you contact the Agency to inquire about regulatory pathways, regulatory requirements, and recommendations about nonclinical and clinical data.

3. Rationale

This guidance makes recommendations as to how you should design and conduct your clinical performance assessment studies (i.e., well-controlled clinical investigations) for your CADe device. These studies may be part of your premarket submission to FDA.[5] The recommendations in this document are meant to guide you as you develop and test your CADe device; they are not meant to specify the full content or type of premarket submission that may be applicable to your device.[6] If you would like the Agency's advice about the classification and the regulatory

[5] This submission may be a premarket notification (510(k)), an application for premarket approval (PMA), an application for a product development protocol (PDP), an application for a humanitarian device exemption (HDE), or an application for an investigational device exemption (IDE).

[6] A 510(k) submission and a PMA application are the most common submission types for the CADe devices addressed in this guidance. As described in the guidance **Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions** (http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm187249.htm), some CADe devices are Class II regulated under 21 CFR 892.2050 and require a 510(k). Other CADe devices are Class III and require a PMA. For more information on the various device classes, see Section 513(a)(1) of the Federal Food, Drug, and Cosmetic Act (the FD&C Act) (21 U.S.C. 360c(a)(1)).

requirements that may be applicable to your device, you may submit a request under Section 513(g) of the Federal Food, Drug, and Cosmetic Act (the FD&C Act).[7]

Regardless of the type of premarket submission you are required to submit for your device, we recommend that you request the Agency's review of your protocols prior to initiating your standalone performance assessment and clinical performance assessment studies for your CADe device. To request the Agency's review of your protocols, you may submit a pre-submission to the Agency.

4. Clinical Study Design

The clinical performance assessment of a CADe device is intended to demonstrate the clinical safety and effectiveness of your device for its intended use, when used by the intended user and in accordance with its proposed labeling and instructions.

As described above in the scope, a CADe device, by design, is intended to identify data that may reveal abnormalities during interpretation of patient radiology images or data by the clinician. There is a complex relationship between the CADe output and the clinician such that clinical performance may depend on a variety of factors that should be considered in any study design including:

- timing of CADe application in the interpretive process (e.g., concurrent or second read);
- physical characteristics of the CADe mark, i.e., size and shape, type of boundary (e.g., solid, dashed, circle, isocontour), and proximity of the CADe mark to the abnormality;
- user's knowledge of the type of abnormalities that the CADe is designed to mark; and
- number of CADe marks.

Your clinical performance assessment should use a well-controlled study design to preclude or limit various biases that might impact conclusions on the device's safety or effectiveness (see **Section 9. Appendix**). We especially recommend this when your clinical performance assessment is performed in a laboratory setting (i.e., off site of the clinical arena) since it is difficult to duplicate the clinical environment in the laboratory. Some study designs that may be utilized to assess your CADe device include:

- A field test or prospective reader study (e.g., randomized controlled trial) that evaluates a device in actual clinical conditions. A field test may not be practical in situations, for example, where there is very low disease prevalence that may necessitate enrollment of an excessively large number of patients.
- A retrospective reader study consisting of a retrospective case collection enriched with diseased/abnormal cases is a possible surrogate for a field test.
- A stress test is another option for the clinical performance assessment of some CADe devices. A stress test is a retrospective study enriched with patient cases that contain more challenging imaging findings (or other image data) than normally seen in routine clinical practice but that still fall within the device's intended use population (see **Section**

[7] Section 513(g) of the FD&C Act (21 U.S.C. 360c(g)) provides a means for obtaining the Agency's views about the classification and the regulatory requirements that may be applicable to your device.

5. Study Population). Note that the use of sample enrichment will likely alter reader performance in the trial compared with clinical practice because of the differences in disease prevalence (and case difficulty for stress testing) between the trial and clinical practice.

The clinical performance assessment of CADe devices is typically performed by utilizing a multiple reader multiple case (MRMC) study design, where a set of clinical readers (i.e., clinicians evaluating the radiological images or data in the MRMC study) evaluate image data under multiple reading conditions or modalities (e.g., readers unaided versus readers aided by CADe). The MRMC design can be "fully-crossed" whereby all readers independently read all of the cases. This design offers the greatest statistical power for a given number of cases. However, non-fully crossed study designs may be acceptable, for example in prospective studies where interpretations of the same patient data by multiple clinicians may not be feasible.

Whether you decide on a fully-crossed study design or not, we recommend the use of an MRMC evaluation paradigm to assess the clinical performance of a CADe device using one of the study designs described above. A complete clinical study design protocol should be included in your submission. Pre-specification of the statistical analysis is a key factor for obtaining consistent and convincing scientific evidence. We recommend you provide:

- a description of the study design;
- a description of how the imaging data are to be collected (e.g., make and model of the imaging device imaging protocol);
- a copy of the protocol, including the following:
    - hypothesis to be tested and study endpoints,
    - plans for checking any assumptions required to validate the tests,
    - alternative procedures/tests to be used if the required assumptions are not met,
    - study success criteria that indicate which hypotheses should be met in order for the clinical study to be considered a success,
    - statistical and clinical justification of the selected case sample size,
    - statistical and clinical justification of the selected number of readers,
    - image interpretation methodology and relationship to clinical practice,
    - randomization methods, and
    - reader task including rating scale used (see **Subsection 4.4**. **Rating Scale**);
- the reader qualifications and experience;
- a description of the reader training;
- a statistical analysis plan (i.e., endpoints, statistical methods) with description of:
    - the process for defining truth (see **Section 6. Reference Standard**),
    - the details of the scoring technique used (see **Subsection 4.5**. **Scoring**), and
    - any results from a pilot study supporting the proposed design.

Valid estimation of clinical performance for CADe devices is dependent upon sound study design. Aspects of sound clinical study design should include the following:

- study populations (both diseased and normal cases) are appropriately representative of or sampled from the intended use population;

- study design avoids confounding of the CADe effect, e.g., reading session effects;
- sample size is sufficient to demonstrate performance claims;
- truth definition is appropriate for assessment of performance, and uncertainty in the reference standard is correctly accounted for in the study analysis, if applicable;
- appropriate data cohorts are represented in the data set;
- readers are selected such that they are representative of the intended population of clinical users; and
- imaging hardware are selected such that they are consistent with current clinical practice.

4.1 Evaluation Paradigm and Study Endpoints

You should select study endpoints to demonstrate that your CADe device is effective (i.e., 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).[8] Selection of the primary and secondary endpoints will depend on the intended use of your device and should be fixed prior to initiating your evaluation. Performance metrics based on the receiver operating characteristic (ROC) curve or variant of ROC (e.g., free-response receiver operating characteristic (FROC) curve or location-specific receiver operating characteristic (LROC) curve), in addition to sensitivity (Se) and specificity (Sp) at a clinical action point will be likely candidates as endpoints. An ROC based endpoint allows evaluation of the device over a range of operating points. An ROC curve is a plot of Se versus 1-Sp and is a summary of diagnostic performance of a device or a clinician. An FROC curve is a plot of Se versus the number of false positive marks per image set. FROC metrics summarize diagnostic performance when disease location and multiple disease sites per patient are accounted for in the analysis. See Wagner, *et* al.[9] and the IRCU Report 79 [10] for additional details on these assessment paradigms. Reporting the Se and Sp pair allows for the evaluation of the device at a clinical threshold or cut-point the reader would act upon. Se is defined as the probability that a test is positive for a population of patients with the disease/condition/abnormality while Sp is defined as the probability that the test is negative for a population of normal patients (i.e., patients without the disease/condition/abnormality).[11]

Se and Sp estimates should be based on an explicit clinical determination by the clinical readers (e.g., recall or no recall), not indirectly from ratings data used to generate the ROC curves. Data collection for ROC, Se and Sp can be done simultaneously within a single reader study. An example reading scenario could be to first have the readers give a clinically justified binary response for Se and Sp evaluation and then to immediately follow with a rating response consistent with the binary response to be used for ROC evaluation.

---

[8] See 21 CFR 860.7(e).

[9] Wagner, R. F., Metz, C. E., and Campbell, G., "Assessment of medical imaging systems and computer aids: A tutorial review," **Acad. Radiol.** 14:723–48, 2007.

[10] ICRU Report 79, "Receiver Operating Characteristic Analysis in Medical Imaging," Vol.8 No.1 (2008), Oxford University Press (ISSN 1473-6691).

[11] Altman, D.G., **Practical Statistics for Medical Research**, Boca Raton, Chapman & Fall/CRC, 1991.

You may employ various summary performance metrics to assess the effectiveness of the use of your CADe device by readers (and such metrics may vary based on the specific device and clinical indication). Examples of these include:

- area, partial area, or other measures of the ROC curve,
- area, partial area, or other measures of the FROC curve,
- area, partial area, or other measures of the LROC curve,
- reader Se and Sp (or Se and recall rate[12]), and
- reader localization accuracy.

We recommend using an ROC summary performance metric as part of your primary analysis, although we recognize that alternate performance metrics may also be appropriate. As mentioned above, we recommend that you include Se and Sp as a secondary endpoint in your analysis when an ROC summary performance metric is used. Reporting Se and Sp (or Se and recall rate) may provide additional information for understanding the expected impact of a device on clinical practice. We also recommend you contact the Agency when you are considering an alternate performance metric.

For study endpoints based on the area under the ROC/FROC/LROC curve or partial area under the ROC/FROC/LROC curve, we recommend that you provide plots of the actual curves along with summary performance information for both parametric and non-parametric analysis approaches when possible. See Gur *et* al.[13] for potential limitations of relying on only one type of ROC analyses. Finally, you should check all methods utilized in your analysis for the adequacy of their fit to the data as appropriate.

The selection of lesion-based, patient-based, or another unit-based measure of performance as a primary or secondary endpoint will depend on the intended use and the expected impact of the device on clinical practice. Powering any additional units-based analyses for statistical significance should not be necessary unless you intend to make specific performance claims.

We recommend that you describe your statistical evaluation methodology, and provide results including:

- overall reader performance;
- stratified performance by relevant confounders or effect modifiers (e.g., lesion type, lesion size, lesion location, scanning protocol, imaging hardware, concomitant diseases) (see **Section 5. Study Population**) (powering each cohort for statistical significance is not necessary unless you are making specific subset performance claims); and
- confidence intervals (CIs) that account for reader variability, case variability, and truth variability or other sources of variability when appropriate.

---

[12] Recall rate refers to the percentage of patients (including diseased and non-diseased patients) that are called back or recalled for additional medical assessment.

[13] Gur, D., Bandos, A.I., and Rockette, H.E., "Comparing Areas under Receiver Operating Characteristic Curves: Potential Impact of the Last Experimentally Measured Operating Point," **Radiology** 247:12–15, 2008.

We recommend that you identify and validate your analysis software.[14] You should provide a reference to the analysis approach used, clarify the software implementation, and specify a version number if appropriate. Certain validated MRMC analysis approaches, examples of which can be found in the literature or obtained online, may be appropriate for your device evaluation depending on its intended use and conditions of use.[15,16]

The definitions of a true positive, true negative, false positive, and false negative CADe mark should be consistent with the intended use of the device and the characterization of the reference standard (see **Section 6**. **Reference Standard**).

4.2 Control Arm

We recommend you assess the clinical performance of your CADe device relative to a control modality. For PMA submissions, a study control arm that uses conventional clinical interpretation (i.e., interpretation without the CADe device) should generally be the most relevant comparator in CADe performance assessment. For CADe devices intended as second readers, another possible control is double reading by two clinicians. For 510(k) submissions, these controls or a direct comparison with the predicate CADe device should generally be appropriate for establishing substantial equivalence. Other control arms may be valid. We recommend you contact the Agency to discuss your choice of a control arm prior to conducting your clinical study.

The study control arm should utilize the same reading methodology as the device arm and be consistent with clinical practice. The same population of cases, if not the same cases themselves, should be in all study arms to minimize potential bias. For designs that include distinct cases in each study arm, we recommend you provide a description and flow chart demonstrating how you randomized patients and readers into the different arms.

---

[14] For more information on MRMC analysis software, see, for example, Obuchowski, N. A., Beiden, S. V., Berbaum, K. S., Hillis, S. L., Ishwaran, H., Song, H. H., and Wagner, R. F., "Multi-reader, multi-case ROC analysis: An empirical comparison of five methods," **Acad. Radiol.** 11: 980–995, 2004.

[15] For MRMC literature references, see, for example: Metz, C. E., "Fundamental ROC analysis," **Handbook of Medical Imaging.** Vol. 1. Physics and Psychophysics. Beutel J, Kundel HL, and VanMetter RL (Eds.) SPIE Press, 751–769, 2000; Wagner, R. F., Metz, C. E., and Campbell, G., "Assessment of medical imaging systems and computer aids: A tutorial review," **Acad. Radiol.** 14:723–48, 2007; Obuchowski, N. A., Beiden, S. V., Berbaum, K. S., Hillis, S. L., Ishwaran, H., Song, H. H., and Wagner, R. F., "Multi-reader, multi-case ROC analysis: An empirical comparison of five methods," **Acad. Radiol.** 11: 980–995, 2004.

[16] For online access to software that analyzes MRMC data based on validated techniques, see, for example: LABMRMC software and general ROC software, The University of Chicago: http://metz-roc.uchicago.edu/ (for either quasi-continuous or categorical data); University of Iowa MRMC software: http://perception.radiology.uiowa.eduhttp://perception.radiology.uiowa.edu (for categorical data); OBUMRM software: http://www.bio.ri.ccf.org/html/obumrm.html.

4.3 Reading Scenarios and Randomization

Reading scenarios in the clinical evaluation should be consistent with the intended use of the device. The following are examples of reading scenarios that may be part of a CADe clinical evaluation[17].

- Devices for second reader use only (sequential design):
  - a conventional reading without the CADe device (i.e., reader alone);
  - a second-read in which the CADe output is displayed immediately after conducting a conventional interpretation (this reading could occur within the same reading session as the conventional read in what is termed a "sequential" reading scheme).
- Devices for concurrent reading (cross-over design):
  - a conventional reading without the CADe device (i.e., reader alone);
  - a concurrent or simultaneous read in which the CADe output is available at any time during the interpretation process (this reading would be conducted in a separate reading session from the conventional read).

You should randomize readers, cases, and reading scenarios to reduce bias in performance measures. We recommend you describe your randomization methodology and provide an associated flowchart. One approach to randomization is to make use of the principle of Latin squares as part of a block design to the reader study.[18]

In case of multiple reading sessions where the same cases are read multiple times, we recommend that you separate each reading session in time by at least four weeks to avoid memory bias. However, longer time gaps may be advisable. For shorter or longer time gaps between reading sessions, we recommend you provide data supporting your proposed time gaps.

4.4 Rating Scale

You should use conventional medical interpretation and reporting for lesion location, extent, and patient management. ROC-based endpoints (see **Subsection 4.1**.**Evaluation Paradigm and Study Endpoints**) may support collecting data with a finer rating scale (e.g., a 7-point or 100-point scale) when readers rate the lesion and/or disease status in a patient. We recommend providing training to the readers on the use of the rating scale (see **Subsection 4.6**. **Training of Study Participants**).

4.5 Scoring

We refer to the procedure for determining the correspondence between the reader's interpretation and the ground truth (e.g., disease status) as the scoring process. The scoring process and the scoring definition are important components in the clinical assessment of a CADe device and you should describe them. We recommend you describe the process (i.e., rationale, definition, and criteria) for determining whether a reader's interpretation

---

[17] Obuchowski NA, Meziane M, Dachman AH, Lieber ML, Mazzone PJ. What's the control in studies measuring the effect of computer-aided detection (CAD) on observer performance? **Acad Radiol.** 2010;17(6):761-7.

[18] Bailey. 2008. *Design of Comparative Experiments*. Cambridge University Press.

corresponds to the truth status established during the truthing process (see **Section 6. Reference Standard** for information on the truthing process).

In this document, we describe scoring in terms of the clinical performance assessment. A different type of scoring is used to evaluate device standalone performance, which is described in the guidance entitled **Computer-Assisted Detection Devices Applied to Radiology Images and Radiology Device Data - Premarket Notification [510(k)] Submissions.**[19]

The scoring process for the clinical studies should be consistent with the abnormalities marked by the CADe and the intended use of your device. You should describe and fix the scoring process prior to initiating your evaluation. In your description of the scoring process, we recommend you indicate whether the scoring is based on:

- electronic or non-electronic means;
- physical overlap of the boundary, area, or volume of a reader mark in relation to the boundary, area, or volume of reference standard;
- relationship of the centroid of a reader mark to the boundary or spatial location of reference standard;
- relationship of the centroid of the reference standard to the boundary or spatial location of a reader mark;
- interpretation by reviewing reader(s); or
- other methods.

For scoring that relies on interpretations by reviewing readers, we recommend you provide the number of readers involved, their qualifications, their levels of experience and expertise, the specific instructions conveyed to them prior to their participation in the scoring process, and any specific criteria used as part of the scoring process. When multiple readers are involved in scoring, you should describe the process by which you combine their interpretations to make an overall scoring determination or how you incorporate their interpretations into the performance evaluation, including how you address any inconsistencies.

4.6 Training of Clinical Readers

We recommend you specify instructions and provide training to clinical readers in the study on the use of the CADe device and the details on how to participate in the clinical study. Training should include a description of the device and instructions for how to use the device. For specialized reading instructions or rules (e.g., rules for changing initial without-CADe interpretation when reviewing the CADe marks), we recommend you justify their clinical relevance according to reading task, clinical workflow, and medical practice.

We also recommend that you provide training to the readers on the use of the rating scale (see **Subsection 4.4. Rating Scale**), especially if such a rating scale is not generally utilized in clinical practice. Such training helps avoid incorrect or un-interpretable results. We

[19] http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm187277.htm

recommend that reader training include rating a representative set of normal and abnormal cases according to the study design methodology, and making use of cases that are not part of the testing database.

5. Study Population

You may collect patient data (i.e., cases) prospectively or retrospectively based on well-defined inclusion and exclusion criteria. We recommend that you provide the protocol for your case collections. Note that cases collected for your clinical trial should be independent of the cases used during your device development. An acceptable approach for acquiring data is the collection of consecutive cases that are within the inclusion and outside of the exclusion criteria from each participating collection site.

Enrichment with diseased/abnormal cases is permissible for an efficient and less burdensome representative case data set. Enrichment may affect reader performance, so the extent of enrichment should be weighed against the introduction of biases into the study design. You may also enrich the study population with patient cases that contain imaging findings (or other image data) that are challenging to clinicians but that still fall within the device's intended use population. This enrichment is often referred to as stress testing and will introduce biases into the study design. Therefore, not all approaches in developing a stress population may be appropriate for use in your clinical assessment. For example, a study population enriched with cases containing small colonic polyps may be appropriate when assessing a general CADe device designed to assist in detecting a wide range of polyp sizes. However, we do not recommend enrichment with cases based on the performance of the CADe device.

You should sample the study population from the intended use population for your device so that the case set includes an appropriate range of diseased/abnormal and normal cases. Selective sampling from cohorts within the intended use population is appropriate when you utilize stress testing. The sample size of the study should be large enough such that the study has adequate power to detect with statistical significance your proposed performance claims. When appropriate, the study should contain a sufficient number of cases from important cohorts (e.g., subsets defined by clinically relevant confounders, effect modifiers, and concomitant diseases) such that clinical performance estimates and confidence intervals can be obtained for these individual subsets. Powering each cohort for statistical significance should not be necessary unless you are making specific subset performance claims. When you make multiple performance claims, a pre-specified statistical adjustment for the testing of multiple subsets would be necessary.

When describing your study population, we recommend you provide specific information, where appropriate, including:

- the patient demographic data (e.g., age, ethnicity, race);
- the patient medical history relevant to the CADe application;
- the patient disease state and indications for the radiologic test;
- the conditions of radiologic testing, e.g., technique (including whether the test was performed with/without contrast, contrast type and dose per patient, patient body mass index, radiation exposure, T-weighting for MRI images) and views taken;

- a description of how the imaging data were collected (e.g., make and model of imaging devices and the imaging protocol);
- the collection sites;
- the processing sites if applicable (e.g., patient data digitization);
- the number of cases:
    - the number of diseased cases,
    - the number of normal cases,
    - methods used to determine disease status, location and extent (see **Section 6. Reference Standard**);
- the case distributions stratified by relevant confounders or effect modifiers, such as lesion type (e.g., hyperplastic vs. adenomatous colonic polyps), lesion size, lesion location, disease stage, organ characteristics (e.g., breast composition), concomitant diseases, imaging hardware (e.g., makes and models), imaging or scanning protocols, collection sites, and processing sites (if applicable); and
- a comparison of the clinical, imaging, and pathologic characteristics of the patient data compared to the target population.

5.1 Data Poolability

Premarket approval applications based solely on foreign clinical data and otherwise meeting the criteria for approval may be approved if, among other requirements, the foreign data are applicable to the United States (U.S.) population and U.S. medical practice and clinical investigators of recognized competence have performed the studies (21 CFR 814.15). You should justify why non-U.S. data reflect what is expected for a U.S. population with respect to disease occurrence, characteristics, practice of medicine, and clinician competency. In accordance with good clinical study design, you should justify, both statistically and clinically, the poolability of data from multiple sites. We recommend that premarket notification submissions follow similar quality data practices with regard to foreign data and data poolability. You are encouraged to contact the Agency if you intend to rely on foreign clinical data as the basis of your premarket submission.

5.2 Test Data Reuse

The Agency recognizes the difficulty in acquiring data for CADe assessment. The Agency also recognizes that the readers and CADe algorithm may implicitly or explicitly become tuned to the test data if the same test set is used multiple times. If no algorithm training has occurred between two tests (e.g., if the sponsor is investigating the effect of a new prompt type), then a new set of readers with no knowledge of the results of the previous study may address concerns about the use of the same test data set. If CADe algorithm training has occurred between tests, or if this is a new CADe algorithm, the Agency encourages evaluation of the CADe system using newly acquired independent test cases (i.e, the Agency discourages repeated use of test data in evaluating device performance). If you are considering data reuse in the evaluation of your CADe device, you should demonstrate that reusing any part of the test data does not introduce unreasonable bias into estimates of CADe performance and that test data integrity is maintained.

In the event that you would like the Agency to consider the reuse of any test data, you should control the access of your staff to performance results for the test subpopulations and

individual cases. It may therefore be necessary for you to set up a "firewall" to ensure those outside of the regulatory assessment team (e.g., algorithm developers) are completely insulated from knowledge of the radiology images and radiological data, and anything but the appropriate summary performance results. You should maintain test data integrity throughout the lifecycle of the product. This is analogous to data integrity involving clinical trial data monitoring committees and the use of a "firewall" to insulate personnel responsible for proposing interim protocol changes from knowledge of interim comparative results. For more information on data integrity for clinical trial data monitoring committees, refer to Section 4.4.1.4 of the guidance entitled **Establishment and Operation of Clinical Trial Data Monitoring Committees for Clinical Trial Sponsors.**[20]

To minimize the risk of tuning to the test data and to maintain data integrity, we recommend you develop an audit trail and implement the following controls when you contemplate the reuse of any test data:

- you randomly select the data from a larger database that grows over time;
- you retire data acquired with outdated image acquisition protocols or equipment that no longer represents current practice;
- you place a small fixed limit on the number of times a case can be used for assessment;
- you maintain a firewall such that data access is tightly controlled to ensure that nobody outside of the regulatory assessment team, especially anyone associated with algorithm development, has access to the data (i.e., only summary performance results are reported outside of the assessment team);
- you maintain a data access log to track each time the data is accessed, including a record of who accessed the data, the test conditions, and the summary performance results; and
- you use new readers in each new clinical reader study.

The purposes of the audit trail include: (1) establishing that you defined the cases in the training and test sets appropriately such that data leakage between training and test sets did not occur; (2) ensuring that you fixed the new CADe algorithm in advance (i.e., before application to the test set); and (3) providing information concerning the extent to which you used the same test set or a subset thereof for testing other CADe algorithms or designs, including results reported to the Agency as well as non-reported results. The controls we are recommending are intended to substantially reduce the chance that you evaluate a new CADe algorithm in a subsequent study using the same test data set that you used for a prior CADe algorithm.

If you reuse test data, you should report the test performance for relevant subsets in addition to the overall performance. These subsets include: (1) the portion of the test set for the new CADe algorithm that overlaps with previously used test sets; and (2) the portion of the test set that you have never been used before. Since these subsets will be smaller in size compared to the overall test set for the new CADe algorithm, confidence intervals for the subsets will be wider. However, the trends for the mean performances would be helpful to indicate whether you have tuned the CADe system to previously used portions of the test set.

[20] http://www.fda.gov/RegulatoryInformation/Guidances/ucm127069.htm

6. Reference Standard

For purposes of this document, the reference standard (also often called the "gold standard" or "ground truth" in the imaging community) for patient data indicates whether the disease/condition/abnormality is present and may include such attributes as the extent or location of the disease/condition/abnormality. We refer to the characterization of the reference standard for the patient, e.g., disease status, as the truthing process.

We recommend that you provide the rationale for your truthing process and indicate if it is based on:

- the output from another device;
- an established clinical determination (e.g., biopsy, specific laboratory test);
- a follow-up clinical imaging examination;
- a follow-up medical examination other than imaging; or
- an interpretation by a reviewing clinician(s) (i.e., clinical truther(s)).

We also recommend that you describe the methodology utilized to make this reference standard determination (e.g., based on pathology or based on a standard of care determination). For truthing that relies on the interpretation by a reviewing clinician(s), we recommend you provide:

- the number of clinical truthers involved;
- their qualifications;
- their levels of experience and expertise;
- the instructions conveyed to them prior to participating in the truthing process;
- all available clinical information from the patient utilized by the clinical truthers in the identification of disease/condition/abnormality and in the marking of the location and extent of the disease/condition/abnormality; and
- any specific criteria used as part of the truthing process.

When multiple clinical truthers are involved, you should describe the process by which you combine their interpretations to make an overall reference standard determination and how your process accounts for inconsistencies between clinicians participating in the truthing process (truth variability) (see **Subsection 4.1. Evaluation Paradigm and Study Endpoints**). Note that clinicians participating in the truthing process should not be the same as those who participate in the core clinical performance assessment of the CADe device.

7. Reporting

Reporting of performance results may be guided by the FDA Guidance entitled **Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests; Guidance for Industry and FDA Reviewers**.[21] We recommend submitting electronically the data used in any statistical analysis in your study including the following:

- patient information,

[21] http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm071148.htm

- disease or normal status,
- concomitant diseases,
- lesion size,
- lesion type,
- lesion location,
- disease stage,
- organ characteristics.
- imaging hardware,
- imaging or scanning protocol,
- imaging and data characteristics (e.g., characteristics associated with differences in digitization architectures for a CADe using scanned films),
- and statistical analysis.

For more information on submitting data electronically, please see the FDA document entitled **Clinical Data for Premarket Submissions**.[22]

8. Postmarket Planning for PMAs

The Agency applies the "Total Product Life Cycle (TPLC)" model to promote and protect the public health and may require you to conduct a post-approval study as a condition of approval in a PMA approval order (21 CFR 814.82(a)(2)). FDA determines whether one is necessary on a case-by-case basis.

In the case where a post-approval study is determined to be appropriate for a CADe PMA submission, the PMA application should include a postmarket plan to assess the continued safety, effectiveness, and reliability of an approved device for its intended use. FDA intends to work interactively with you to finalize the postmarket plan and any post-approval study protocols prior to the approval decision so that the study is ready to implement if FDA approves the device.

For additional information, please refer to the FDA Guidance entitled **Procedures for Handling Post-Approval Studies Imposed by PMA Order.**[23]

9. Appendix

9.1 Potential Sources of bias in a retrospective reader study

Despite their practical value, retrospective reader studies for evaluating CADe devices may generate estimates of CADe performance that are subject to one or more potential sources of statistical bias. Statistical bias is a tendency for a performance estimate in a study to be

[22] http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/HowtoMarketYourDevice/PremarketSubmissions/ucm136377.htm

[23] http://www.fda.gov/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/ucm070974.htm

misaligned with the true performance in the intended use population. Many sources of statistical bias can often be minimized or at least mitigated through good study design. Some potential sources of bias in reader studies include:

- Selection Bias: The sample of subjects (or readers) is not representative of the target population.
- Spectrum Bias: The sample of subjects (or readers) studied does not include a complete spectrum of the target population.
- Imperfect Reference (Gold) Standard Bias: The reference procedure is not 100% accurate at classifying subjects by presence or absence of the condition of interest (e.g., breast cancer, polyps in colon).
- Verification Bias: Statistical analysis of diagnostic performance is based only on subjects verified for presence or absence of the condition of interest by the reference standard.
- 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.
- Context Bias: When the sample prevalence of the abnormal condition differs greatly from the prevalence in the target population, the reader's interpretations may be affected, resulting in biased estimates of his/her diagnostic accuracy.

Selection bias is introduced when subjects (cases) selected for study are not representative of the intended use population. Retrospective selection of subjects with available images can introduce a selection bias. Random or consecutive sampling of subjects may eliminate or mitigate selection bias. Selection bias can also occur in terms of the readers selected to participate in the study. Readers should reflect the population of readers that will use the device. A small number of readers having similar training or clinical experiences (e.g., readers all from the same clinic) may not generalize to the full population of readers expected to use a CADe device.

Spectrum bias is a special form of selection bias in which the study subjects represent an incomplete spectrum of characteristics of subjects in the intended use population (i.e., important subgroups are missing from the study).

Enriching a CADe study with subjects that are difficult to diagnose (i.e., stress testing) can bias the CADe effect. For example, if the CADe is most helpful for these difficult cases, then the CADe effect can be enhanced with stress testing relative to the intended use population. Nonetheless, stress testing is encouraged because it provides value in evaluating CADe in important subgroups, and results in study designs with smaller sample sizes when applied appropriately.

Enriching a CADe study with subjects having the abnormal condition produces biased estimates of CADe positive and negative predictive value because these depend on the prevalence of the abnormal condition. "Corrected" estimates of predictive values would have to rely on an estimate of prevalence external to the study. The enhanced prevalence may also indirectly cause biased estimates of area under the ROC curve (AUC), of Se, and of Sp if readers are "tipped off" to the enhanced prevalence in the study and change their reading behavior, accordingly creating context bias.

Retrospective reading of image sets may also introduce bias into CADe performance. The basic concern is that reading behavior may change relative to actual clinical practice because readers know that they are participating in a study in which patient management is not affected by their diagnostic readings.

Ground truth determination for the presence or absence of the condition of interest can be subject to several sources of bias. A reference standard is used to determine ground truth. *Imperfect reference standard bias* refers to the reference standard misclassifying some subjects as having or not having the condition. For example, ground truth determination is incomplete if subjects diagnosed as negative for the abnormal condition on initial evaluation are not followed up for confirmation that they were indeed free of the condition.

Verification bias refers to the ground truth being missing for some subjects. If in the statistical analysis, only the subjects on whom ground truth has been established are included, estimates of CADe performance can be biased.

The study design of a retrospective reader study can produce potential sources of statistical bias. For example, in a sequential reading design, readers are instructed to read sequentially first unassisted by CADe, and then aided by CADe as a second reader. The comparison is between an unaided read and the unaided read combined with the CAD-aided read. The sequential reading design is attractive because the proximity of the readings in the two modalities minimizes intra-reader variability. However, if the reading condition in the indications for use (IFU) of the device are different from that in the reader study, a concern might be that the effect of CADe on the diagnostic performance of readers may be confounded with the effect due to the additional time given readers to read each case. Another concern is that a reader may undercall the unaided reading relative to the reading aided by CADe, producing an enhanced CADe effect. If this concern is real, mitigation may be to randomize a fraction of the cases to be read only in the initial unaided reading mode. Randomization is revealed only after the initial unaided read is made. Randomization of the cases should be done separately for each reader in a way that ensures that all cases are given a CADe aided reading by some readers.

An alternative to the sequential reading design is the *cross-over design*. Cases are read unaided and aided by the CADe output in two independent reading sessions separated by a washout period to erase reader memory of the images. Half of the cases are randomly assigned to group A and the other half to group B. In reading session 1, group A cases are read unaided and group B cases are read aided by CADe. In reading session 2, the cases are "crossed over" to the other modality (aided reading for group A, unaided reading for group B). Any effects the particular reading sessions have on the readings cancel in the comparison of the two modalities. However, relative to the sequential reading design, the two reading sessions contribute additional variability to estimate of the CADe effect. The cross-over design may be particularly appropriate for concurrent reading. It can also be generalized for use in evaluating more than two modalities (e.g., if in addition to unaided reading, the CADe has two or more modalities itself).

References

1. Pepe, 2003, *Statistical evaluation of medical tests for classification and prediction*, Oxford Press.

2. Zhou, Obuchowski, McClish, 2002, *Statistical Methods in Diagnostic Medicine*, Wiley: New York.