Provisional Translation (as of June 2024)

This English version of the Japanese review point is provided for reference purposes only. In the event of any inconsistency between the Japanese original and the English translation, the former shall prevail.

Review Point for Supporting Software for Detecting Lesion with Endoscopic Imaging

Pharmaceuticals and Medical Devices Agency (PMDA)
March 7, 2023

TABLE OF CONTENTS

Review Points for Supporting Software for Detecting Lesion with Endoscopic Imaging

Introduction

At the time of making an approval application, information on the review point will be organized and released from medical device programs that have been approved after 2014.

- This review point shall indicate necessary endpoints, etc. for medical devices shown in the specified scope to contribute to the improvement of efficiency in preparation of materials and acceleration of reviews for an approval application.
- This review point shows the concept of review based on the current scientific knowledge, and it shall be reviewed and revised as needed according to future advances in science and technology.

1. Products Covered by This Document

According to "Release of Next-Generation Medical Devices Evaluation Criteria" (PSEHB/MDE Notification No. 0523-2 dated May 23, 2019) Attachment 4 "Evaluation Criteria for Medical Diagnostic Imaging Support Systems Using Artificial Intelligence Technology," Computer-Aided Diagnosis is defined as follows.

CADe (Computer-Aided Detection): Standalone software with a function of automatically detecting regions with suspected lesions on an image by a computer and marking their regions or a device incorporating such software. A computer processes medical image data alone or both medical image data and laboratory data, and supports detection of lesions or abnormal values.

CADx (Computer-Aided Diagnosis): Standalone software with a function of outputting quantitative data such as differential diagnosis of benign and malignant lesions and disease progression as numerical values, graphs, etc. in addition to detection of regions with suspected lesions or a device incorporating such software. This includes those that provide diagnostic support by providing candidate diagnostic results, information on risk assessment, etc.

This document summarizes the review points for an approval application for CADe, which provides diagnostic support during endoscopy. It does not suggest severity of the disease or classify detected findings (those that detect multiple types of findings are included, but those that classify and present the types of detected findings are excluded). In other words, this review point applies to the supporting software for detecting lesion with endoscopic imaging specified in No. 1991 of the Appendix 2 of specially controlled medical devices, controlled medical devices, and general medical devices designated by the Minister of Health, Labour and Welfare according to the stipulations in Article 2, Paragraphs 5 to 7 of the Act on Securing Quality, Efficacy and Safety of Products Including Pharmaceuticals and Medical Devices (MHLW Ministerial Announcement No. 298 of 2004). It also applies to those used as concurrent readers.

- This document describes:
  - Product submitted for registration: A product intended for approval. It is intended for the product submitted for registration in an approval application, and the product submitted for face-to-face consultation, etc.
  - Authorities: It means regulatory authorities. It refers to PMDA at the time of review/consultation.
  - First Reader: CAD frst performs an interpretation process alone, and the physician’s interpretation is restricted to only the CAD-marked images.
  - Concurrent Reader: CAD and a physician perform an interpretation at the same time.
  - Second Reader: A physician frst interprets images without CAD, and then CAD performs an interpretation.

2. Description of the Product Submitted for Registration

2.1 Organization of the role in clinical practice

In order to discuss the sufficiency of the evaluation package required for the product submitted for registration and the validity of the evaluation test, it is important to clarify not only the specifications (input data, output data, numerical values related to performance, etc.) of the product submitted for registration but also the role in clinical practice (hereinafter referred to as "ROLE") of it and share it with the authorities. The ROLE refers to by whom and for what purpose the product submitted for registration is used in clinical practice.

Even for CADe intended to support endoscopic diagnostic imaging, it is important to organize the ROLE. For example, the following variations can be imagined for CADe that supports detection of medical image findings (not limited to this variation).

Example)

- Use of CADe will help prevent overlooking during endoscopy. (e.g., to support endoscopists in preventing overlooking, to raise the level of diagnostic technique for endoscopy among physicians with little experience of endoscopic diagnosis, and to equalize the diagnostic performance at each medical institution)

In each of the above, the performance that the product submitted for registration should aim for and the test methods differ. To understand the ROLE, it is better to organize information such as the following: However, the explanation should be added appropriately according to the development concept of the product submitted for registration.

- What is the target disease, and who are the patients, etc.? (e.g., patients with colorectal lesions (e.g., colorectal polyp, precancerous colorectal cancer, or early colorectal cancer))
- Who (specialists, non-specialists, etc.) will be allowed to use the product and how it will be used? (e.g., endoscopist, all endoscopy physicians)
- What are the challenges in the current medical practice for the target disease/patients, etc.?
  - If there is information that can be explained quantitatively from guidelines, literature, etc., it is desirable to explain it together with such information. (e.g., prevalence and sensitivity/specificity etc. of endoscopy performed to check the presence or absence of colorectal lesions)
- How will the product submitted for registration solve the problem?
  - What and to what extent can it detect clinically and can it be a support?
- How will the existing medical practice be changed by introducing the product submitted for registration to the medical practice?
  - How does it affect the flow of diagnosis and treatment guidelines?
  - How will it affect physicians and institutions using the product submitted for registration? (Advantages and disadvantages)

    Advantages and disadvantages of patients diagnosed or treated using the product submitted for registration.

- Whether it is possible to implement risk management measures so that the above-mentioned the advantages outweigh the disadvantages.

How existing medical practice is performed and how the product submitted for registration is introduced are important not only for performance evaluation but also for examination of risks of the product (mainly impact of false-positive/false-negative). It is desirable to explain existing medical practice with reference to various practice guidelines, etc.

At the time of consultation, it is desirable to also explain the proposed intended use to obtain approval based on the above (*).

* It is only intended to promote the description to understand what kind of approval the applicant wants to obtain at the time of consultation. It should be noted that the final intended use will not be determined at the time of consultation but will be determined after review.

2.2 Design Concept

Based on the ROLE explained in Section 2.1, it is necessary to organize the functions/performance (design concept) for which the product was developed. In other words, it may be an explanation of what functions were considered necessary to achieve the ROLE and how much performance was considered necessary for the functions.

Items in Sections 2.2.1 to 2.2.3 should be organized based on the design concept of CADe intended to support endoscopic diagnostic imaging.

2.2.1 Function

It is necessary to explain the specific functions of the product submitted for registration. The functions and specifications need to be concretized at the time of an approval application, but it is inevitable that there are partially uncertain functions/specifications depending on the development phase at the time of seeking consultation. In this case, it should be described to make it clear that these specifications are under consideration.

The following are examples of elements that describe the CAD function of the product submitted for registration. It should be noted that specific descriptions are required according to the functions, etc. specific to each product submitted for registration.

* Descriptions such as "function to perform ●●, etc." and "function to process ○ ○" alone cannot make us understand the whole picture of the functions of the product submitted for registration and the output functions. Please pay attention when preparing application materials/consultation materials so that they can be described comprehensively and concretely.

* Because reviews/consultations are conducted based on application/consultation materials, reviewers cannot recognize functions without descriptions. If new functions are recognized during the review/consultation, it may be difficult to extend the review/consultation or continue the review. Therefore, please prepare to be able to share all the functions of the product submitted for registration with reviewers.

(1) Input

- What is the subject population to be analyzed? (e.g., patients suspected of having colorectal lesions (e.g., colorectal polyp, precancerous colorectal lesion, or early colorectal cancer, persons undergoing disease screening))
- What is the imaging modality? (e.g., flexible endoscope, rigid endoscope)
- What are the imaging conditions? (e.g., identification of image mode (white light mode, image enhancement mode, magnifying endoscopic image, etc.))
- What is the vendor of the imaging modality? (e.g., marketing authorization holder of endoscope, image processing device, etc. used)
- Is the image with image processing (enhancement processing, reconstruction processing, etc.) used/or are raw data before image processing used?
- Presence/absence of dye use

* It should be set as an input condition to specify that the product submitted for registration can demonstrate appropriate performance. It should also be noted that these conditions should be considered for evaluation during the test.

(2) Detection target

- What kind of findings and properties are included in the target? (e.g., type of polyp/macroscopic type)
- If there are any special conditions such as findings difficult for physicians to find, what are the details? (e.g., lesions that are too small in size, slur, halation, residue, etc.)
- What findings are not detected? (e.g., non-epithelial lesions, inflammatory bowel disease)
- Whether or not the findings, etc. can be detected by physicians from the images to be analyzed in conventional medical practice.

(3) Analysis principle

- What is the analysis principle?
  - If it is designed deductively, what is the processing algorithm?
  - If a model is realized using machine learning, what is the training algorithm and what are the development data (data collection facilities, how to annotate labels, how much the volume of training data)? If there is a decision threshold, what is it? (See also Section 5.)
- How is the final output determined?
  - What parts are combined or points determined and presented?
  - What are the decision thresholds?

(4) Output

- Whether it is displayed on the main display, whether it is essential to use the sub-display and displayed on the sub-display, etc.
- Is the case/image unit flagged? Or will the potential finding to be detected on the image be

flagged?

- What kinds of flags such as points, bounding boxes, circle heat maps, etc. are given in the case of flagging on images? Are the dimensions presented constant or variable (e.g., depending on the size of the finding)?
- Under what conditions is a flag displayed (e.g., a lesion is detected in a ○ frame sequence)
- At what speed is the flag displayed? (e.g., processing speed that becomes real time for the physician, certain delays, etc.)
- How many flags are displayed at the same time? Is there any upper limit for the number of labels? What is the maximum number of flags displayed?
- If multiple types of findings are to be detected, are these types of findings displayed without distinction or with distinction?
- Is the confidence level (or equivalent scale, etc.) calculated in the analysis process displayed? In this case, what is the significance of the confidence level displayed to provide information?

(5) Other

- What functions other than the CAD function does the product have? For other functions, explain the details of the functions referring to the above (1) - (4).

2.2.2 Directions for use

It is necessary to explain the specific directions for use of the product submitted for registration. The directions for use need to be concretized at the time of an approval application, but it is inevitable that there are partially uncertain directions for use depending on the development phase at the time of seeking consultation. In this case, it should be described to make it clear that the directions for use are under consideration.

The following are examples of factors that describe the directions for use of the product submitted for registration. It should be noted that specific descriptions are required according to the ROLE, etc. specific to each product submitted for registration.

(1) User

- What kind of physicians, etc. use it?
  - If the department, proficiency, qualification, etc. are designated, also explain them.
  - If there are any conditions for restrictions on use by physicians, etc., explain them.

* For example, it should be noted that "specialists" should be considered as the user of the product submitted for registration if the use of "specialists" is not denied even if the main user is a "non-specialist."

(2) Directions for use

- How do the concurrent reader and second reader, etc. confirm the results of the product submitted for registration in actual medical practice? (e.g., are results displayed prior to physician diagnosis?)

* Please note that if the reading test described in Section 3.2 are conducted, the test design will be considered based on the directions for use. It should also be noted that the directions for use need to be examined in advance when the validity of the test design is examined.

2.2.3 Performance

Explain how much performance, you considered, the CAD function of the product submitted for registration should have for what kind of images based on the design concept. "What kind of images" generally corresponds to Section 3.2.1 (1). “How much performance" includes, for example: It should be noted that this does not necessarily match the evaluation design of the single performance test (it is acceptable to evaluate the entire application package).

- The target of detection can be detected with the results equivalent to the diagnostic performance of specialists.
- The target of detection can be detected with the results exceeding the diagnostic performance of non-specialists.

2.3 Information on similar products

If there is a similar product to the product submitted for registration, provide information on the product submitted for registration while comparing with the similar product. For how to compare similar products, refer to “Points to consider for preparation of data and documents to be attached to the marketing approval application form of medical devices" (PFSB/MDRMPED Notification No. 0120-9, dated January 20, 2015 by the Counsellor of Minister’s Secretariat, Ministry of Health, Labour and Welfare (the Director of the Medical Device and Regenerative Medicine Evaluation Division)), etc.

3. Evaluation Package

Based on the ROLE and design concept summarized in Section 2, the evaluation package should be considered so that the clinical utility, clinical performance, and basic performance of the product submitted for registration can be evaluated.

High-level conceptual requirements for many CADes are shown below. However, even in CADe in which the same target is detected from the same modality, the details of the test protocol will change depending on the ROLE, display method, etc. such as the intended use, who uses it, how to use it, etc. Therefore, it is useful to refer to what test protocol has been used to evaluate similar precedent products, but it should be considered that basically it is necessary to adjust for the contents of each product submitted for registration.

3.1 High-level conceptual requirements

(* Adjust according to the characteristics of the product submitted for registration.)

(1) The use of the results of analysis by the product submitted for registration for the intended input data shall improve the diagnostic performance of the intended user. (Clinical utility)

(2) The product submitted for registration shall have clinically significant detection performance

against the intended input data. (Clinical performance)

(3) Processing shall be able to be completed within a clinically acceptable time frame. (Basic performance)

(4) Other functions shall operate as intended. (Basic performance)

(5) The software life cycle shall be appropriately controlled.

As described above, how to evaluate each item will be considered according to the ROLE of these products submitted for registration. The basic idea is that each function of the product submitted for registration needs to be evaluated to a certain extent. However, which function should be evaluated as a verification test, performed as a secondary test, or merely confirm the operation should be examined by the ROLE of the function and impact on patients against incorrect output.

* The term "to be evaluated" refers not only to the conduct of testing, but also to the results of examination based on the results of testing performed. For example, if clinical utility can be explained without conducting a test by discussing clinical performance results and the contents of literature, etc., it can be said that “clinical utility was evaluated.”

In the next section, examples of evaluation of clinical usefulness and clinical performance are described.

3.2 Tests to evaluate clinical utility

The primary objective of clinical utility evaluation is to evaluate whether the development concept of the product submitted for registration has been achieved. If clinical utilitys is directly evaluated by a test, it shall be conducted as a test to directly evaluate the value of the product submitted for registration after simulating the situation where the product has been introduced clinically. In order to directly evaluate the value of the product submitted for registration as a medical device, the test is very persuasive for application for approval. On the other hand, since it is difficult to reproduce actual clinical practice and the participation of physicians, etc. in the test is necessary, there is a certain cost for implementation. In addition to the tests to evaluate the clinical performance described in the next section, the appropriateness and necessity of conducting such tests should be carefully examined.

As an example of a test to evaluate the clinical utility of the product submitted for registration, there is a test in which the diagnostic performance obtained with the product submitted for registration and that obtained without the product submitted for registration are compared and the superiority of the former is evaluated (hereinafter referred to as "reading test"). In this test, it can be verified that the use of the CAD function of the product submitted for registration contributes to the improvement of diagnostic performance in the situation where actual clinical practice is (ideally) replicated.

The test design is also affected by the directions for use of the product submitted for registration (e.g. concurrent reader, first reader, second reader, etc.). Therefore, in such tests to evaluate the clinical utility of the product submitted for registration, interpretation will be performed as "(1) reading without CAD (normal interpretation), followed by (2) reviewing the results of CAD (reading combined with the product submitted for registration)" (continuous evaluation test), and the usefulness

of the product submitted for registration in actual clinical practice can be evaluated by comparing changes in diagnostic performance in (1) and (2). In this regard, in (1), it is desirable to take into consideration so that the results of interpretation are obtained after taking sufficient time as much as possible (in order to deny the possibility that the diagnostic performance was improved by simply reading for a long time (or reading multiple times)).

On the other hand, in the case of the concurrent reader, diagnosis is made with reference to the results of CAD from the first view, and therefore it is necessary to evaluate the improvement of diagnostic performance by combined use of CAD in the first-time patients. If the reading is performed in a continuous evaluation study as exemplified in the second reader, the reading will be performed with CAD while the results at the time of reading without CAD are remembered. Therefore, the results will be different from those in the actual directions for use. It is necessary to eliminate bias (memory bias) by this memory, or devise a test design which is not affected by bias or to evaluate the absence of memory bias retrospectively. In this way, the test design and the test procedure should be designed in consideration of the directions for use.

The test design may also change depending on the output of the product submitted for registration (e.g., adding analysis results to cases, adding analysis results to medical images, or adding analysis results to suspected findings in medical images). If multiple types of findings are detected, it is also necessary to consider whether they are presented distinctively or not. (Points to consider related to this are described in Section 3.5.)

The secondary objective of evaluating clinical utility is to evaluate whether there are false-positive/false-negative results to which physicians using the product are likely to be attracted. Therefore, risk reduction can be expected by taking measures such as provision of information.

See Section 3.3 for points to consider for each component of the test.

< Developmental discussion >

The design of the reading test depends on the development concept and design concept of the product submitted for registration. For example, even if the CAD function of the product submitted for registration is a function to detect some potential findings from medical images, its purpose is diverse, including prevention of overlooking by conventional examination and early detection for early therapeutic intervention. The test design to evaluate whether or not the purpose of this diagnostic support can be fulfilled (e.g., what cases will be evaluated, what kind of users will be evaluated, what will be detected, what will be used as evaluation criteria, what will be meaningful if it can be accomplished) is different. Depending on the ROLE of the product submitted for registration, it may be necessary to conduct a test to see if the patient outcome has been improved by medical practice based on the detected results, rather than simply evaluating whether the target to be detected is more correctly detected. After the development concept of the product submitted for registration is organized, it is necessary to examine how the test design to evaluate the achievement can be assumed and what should be used as endpoints, etc.

3.3 Tests to evaluate clinical performance

The primary purpose of evaluating the clinical performance is to evaluate to what extent the intended output of the product submitted for registration can be performed correctly against input data such as endoscopic images. While the clinical value of the product submitted for registration (whether the diagnostic performance is improved by using the product submitted for registration, etc.) is evaluated in the evaluation of clinical utility, the performance of the product submitted for registration itself will be evaluated in the evaluation of clinical performance. It is mainly the information described in the column of standards related to performance and safety in the application form.

For evaluation of the CAD function, the following two policies are assumed for tests to evaluate clinical performance.

[1] To confirm the efficacy and safety of the product submitted for registration in clinical utility studies. To confirm the clinical performance by the test to specify the performance of the product submitted for registration for which clinical utility has been confirmed.

[2] To confirm the efficacy and safety of the product submitted for registration in clinical performance studies. (A test that evaluates clinical utility should be conducted in the test that evaluates clinical performance.)

In the case of [1] , the clinical utility will be evaluated by conducting a hypothesis verification test, etc., and the efficacy and safety of the product submitted for registration will be confirmed based on the results. In this case, the main purpose of evaluation of clinical performance is to obtain information specifying the performance of the product submitted for registration. On the other hand, if a clinical performance evaluation test that can explain the clinical utility can be conducted, the evaluation strategy in [2] can be considered. For example, if an appropriate performance goal (hereinafter referred to as "PG") which has abundant literature and guidelines related to the ROLE of the product submitted for registration can be explained, the clinical utility and clinical performance may be evaluated in the same test by confirmatory evaluation of the superiority or non-inferiority of the clinical performance of the product submitted for registration to the PG (an appropriate method shall be selected according to the position of the application product). In this case, the validity of PG is an important issue. In particular, it is often possible that there are no published reports, guidelines, etc. completely matched to the ROLE of the product submitted for registration, and therefore it is requested to carefully examine whether an appropriate evaluation system can be designed according to the policy in [2].

The secondary purpose of evaluating the clinical performance is to evaluate the detection performance of the product submitted for registration in rare cases and take measures such as provision of information as necessary. In studies to evaluate clinical utility, it is expected that the data sets are examined and studies are conducted in consideration of the influence on the results due to differences in the patient balance simulating actual clinical practice in addition to a certain variation of patients by the participation of readers, etc.. On the other hand, when evaluating the performance of the product submitted for registration alone among tests to evaluate the clinical performance, even if there is a gap with actual clinical cases, if the design of the product submitted for registration does not take this into

account, there are not any concerns about the patient balance. Rather, it is important to review the data on various variations and to collect information to call attention so that the product will be used more appropriately.

3.4 Other functions

In principle, it is necessary to evaluate all functions of the product submitted for registration. In addition to CAD functions, clinically significant functions and functions that may pose risks to patients due to incorrect operation need to be tested for clinical utility and clinical performance depending on the functions. On the other hand, supplementary functions such as the data input/output function and data storage function can be evaluated to confirm that they can operate as intended.

Next, the conceptual requirements of CADe (hereinafter referred to as endoscopic CADe) intended to aid in diagnosis by detecting candidate regions of colorectal lesions (colorectal polyp, precancerous colorectal lesion or early colorectal cancer, etc.) from images of colonoscopy and alerting physicians to prevent overlooking large intestine lesion candidates based on the results of approval are exemplified.

It should not be included in conceptual requirements for software intended for screening or definitive diagnosis of colorectal lesions and endoscopic CADe learned for improving performance at each medical institution after marketing.

3.5 Conceptual requirements for efficacy and performance required for endoscopic CADe

Basic efficacy and performance required for endoscopic CADe are as follows: Based on these, it is necessary to confirm that necessary design verification tests have been performed.

1) The program should reduce overlooking of colorectal lesions during endoscopy in intended users.
2) The program has sufficient detection accuracy for the target colorectal lesions of the program.
3) When the target colorectal lesion of the program is displayed on the endoscopic image, it should be detected at an adequate processing speed and displayed as specified.
4) The display shall be visible to the endoscopist.

3.6 Conceptual requirements for risks assumed by the use of endoscopic CADe

Risks assumed from the directions for use of endoscopic CADe are as follows. It is necessary to confirm that the residual risk is shown to be within the acceptable range as a result of sufficient risk reduction for these assumed risks.

1) Overlooking of lesion candidates due to a decrease in physician's awareness and accuracy of examination caused by overconfidence in the display function of lesion candidates in the program.
2) Overlooking of lesion candidates that could not be identified by the program on the same image because lesion candidates are identified by the program.
3) Overlooking of lesion candidates that could be detected by the conventional examination

technique because the conventional examination technique is changed in association with the use of the program.

4) Overlooking of lesion candidates because the potential lesions that can be pointed out are not displayed due to delay in processing of the program.
5) Increased burden on physicians and patients when the examination time is extended.

3.7 Exemplification of functions required for endoscopic CADe

The functions set based on the conceptual requirements of endoscopic CADe are shown below.

3.7.1 Lesion detection function (detection sensitivity and specificity)

The sensitivity and specificity of the colorectal lesions shall meet the criteria that can explain the clinical usefulness (for example, they can extract regions estimated to be lesions at the same level of lesion detection rate as that of gastrointestinal endoscopists).

It should be shown that the conceptual requirements 3.5 1) and 2) can be met.

The evaluation method is separately shown in 3.2, 3.3, and 4.

3.7.2 Image data entry

The image shall be input on the combined endoscopic device.

It should be shown that the conceptual requirements 3.5 3) and 3.6 4) can be met.

3.7.3 Warning by screen display at detection

When detected, the endoscopic image shall be highlighted.

It should be shown that the conceptual requirements 3.5 4) can be met.

3.7.4 Warning with sound at detection

A warning sound shall be issued at the time of detection.

It should be shown that the conceptual requirements 3.6.4) can be met.

3.7.5 Timing of result update

The results shall be updated every fixed cycle (e.g. every second).

It should be shown that the conceptual requirements 3.5 3) and 3.6 4) and 5) can be met.

3.7.6 Safety function

(1) Processing of images not to be analyzed

Confirmed that analysis will not be performed if there are images not to be analyzed in the images to be analyzed.

(2) Processing to prevent reduced attention due to excessive detection

Call attention by warning sound/screen coloring when an appropriate exclusion criterion is reached to avoid excessive detection.

4. Test Design Considerations

This section describes considerations in designing the details of the test design.

4.1 Test sample

Test results should be considered as an evaluation of the samples for which approval is required (including product version, etc.). If the evaluation sample is different from the final product, clarify the difference between the evaluation sample and the sample to be approved, and then clarify the reason why the test result can be extrapolated.

Particularly, for verification tests, attention should be paid to the point that the product for which the decision threshold for output of the final product has been determined is also evaluated regarding whether it can achieve the verification items. In the development of general diagnostic support products, there is a method to confirm the diagnostic performance by using ROC, etc., and determine the specification with the highest sensitivity/specificity or the decision threshold expected to have the sensitivity/specificity required based on the ROLE. Although such development policy is not denied, it should be noted that the results obtained at this time cannot be said to be the verified results.

4.2 Handling of human-derived data

In general, clinical studies conducted for the purpose of being attached to an approval application are required to be conducted in compliance with the GCP Ordinance. On the other hand, in the case of a test using the data obtained in daily medical practice, handling based on “Handling of Performance Evaluation Tests of Diagnostic Medical Devices Using Existing Medical Image Data without Involvement of Additional Invasiveness or Intervention” (PSEHB/MDE Notification No. 0929-1, dated September 29, 2021. hereinafter referred to as "0929 Notification) can be considered.

In the tests described in 2. (1) of 0929 Notification, CAD mainly uses medical images only. It should be well considered if the test necessary to evaluate the clinical utility of the product submitted for registration can be designed from medical images alone (for example, if the label in the test can be appropriately defined from medical images alone). If this is difficult, implementation of 0929 Notification (2) or as a clinical study should be considered.

Also, in the tests described in 2. (2) of 0929 Notification, medical data linked to medical images (definitive diagnosis results, etc.) can be used. At that time, it is necessary to pay attention to whether or not the evaluation is appropriate as the evaluation population of the product submitted for registration in terms of the fact that the evaluation will be performed in the population with medical data linked to medical images. For example, if the results of biopsy linked to medical images are treated as the label for the test, the evaluation is limited to the population of patients who underwent biopsy. On the other hand, if the product submitted for registration is used also in patients who do not undergo biopsy, there will be differences between the actual clinical population and the evaluation population. It is necessary to examine how this difference affects the evaluation, and if it affects the evaluation, how to deal with it and to design an appropriate evaluation system.

4.3 Variation of test data set

According to the ROLE of the product submitted for registration, the performance of the product, and the purpose of the test, it is necessary to examine the variation of the test data set to be included in the test. The target population of the product submitted for registration in what clinical phase should be organized, and what test data set to be collected should be clarified. It is also necessary to consider the balance of cases, prevalence, etc. in actual clinical practice depending on the purpose of the test. In addition, the variation related to non-clinical elements should be examined.

Factors may be reduced if it can be explained that no examination is necessary from the analysis principle of the product submitted for registration.

They should be organized and what collection plan will be implemented to achieve the target test data set should be explained.

4.4 Label

The label in the test should be prepared by an appropriate method so that the product submitted for registration can be evaluated based on the ROLE of the product submitted for registration. According to the ROLE of the product submitted for registration, what label should be defined (presence or absence of findings in cases, presence or absence of findings in endoscopic images, position of findings in endoscopic images, etc.) should be considered individually in accordance with the specifications of the product submitted for registration. The validity of the label preparation method affects the interpretation of the test results and the evaluability of the quality, efficacy, and safety of the product submitted for registration. Therefore, it is necessary to clarify the method for preparing the label before performing the test and prepare it so that its validity can be explained.

Many CADes, which are the targets of this review, may be intended to detect clinically significant findings that may be present in the input data. In the previous CADes, many of them are intended to support detection of detectable findings by skilled physicians. For such CADe, it is necessary to design the preparation method so that it can be explained that what is detected from the input data by a general skilled physician in Japan is defined as the correct answer label.

For example, when determining the label based on the interpretation of a skilled physician who prepares a label (hereinafter referred to as a "label determining physician"), the following methods can be considered.

- Interpretation will be made by 3 or more label determining physicians in order to eliminate individuality. The process of assigning interpretation is performed independently.
- The label is determined using the majority vote for the obtained interpretation to eliminate bias due to consultation.

In this regard, the impact of bias due to the validity (job type (physician, etc.), expertise, years of experience, etc.) of label determining physicians, inclusion/exclusion criteria used in the process of preparation of labels, etc. should also be considered.

The label determining physician must be independent of the physician participating in the reading test.

The following reference information may be used to annotate the label: In some cases, a label can

be granted only by medical images, etc., but it is necessary to carefully examine whether a medically significant label can be granted without using reference information shown below. When retrospective tests are conducted, they may not be handled as nonclinical studies even if there is no additional invasion or intervention in patients (see 0929 Notification).

- Diagnosis results from other devices or tests, such as different modalities
- Definitive diagnostic results
- Follow-up diagnostic imaging information

Depending on the future development, there may be a product that detects findings, etc. that are difficult even for skilled physicians to detect only from input data by developing with other modalities or definitive diagnostic results. Such products need to be reconsidered from the ROLE of the product and the overall test package beyond how to create the label.

4.5 Matching with Ground Truth

Since matching with ground truth (hereinafter referred to as "GT") in the test affect the success or failure of the test, the definition of GTs should be specified in advance so that objective assessment can be made. In addition, it is necessary to define GTs in the test in consideration of the purpose of the test, ROLE of the product submitted for registration, and output method of the product.

In the output specifications of CAD, output patterns such as (1) giving determination results to analysis cases, (2) giving results to analysis images (videos), and (3) giving results to detection targets on analysis images, are assumed. For (1) and (2), the GTs may be the agreement between the label of the analysis target or analysis image and the output result of the product submitted for registration. On the other hand, for (3), it is necessary to consider the definition of GTs so that it can also be evaluated that the position presented by the product submitted for registration is consistent with the position of the label in the appropriate positional relationship.

When further classifying the output specifications of the product submitted for registration, various specifications are expected, such as (3-1) specifications that present the detection target area by trimming, (3-2) those that present a bounding box containing the entire detection target, (3-3) those that present the center of gravity and center of the detection target area by a point or a circle with a certain size, and (3-4) those that present the heat map based on the confidence analyzed by CAD by superimposing. Although it should be considered individually based on the specifications, etc. of each product submitted for registration, please consider the definition of correct answers by referring to the following points based on the relationship between the correct answer label and output specification.

Table 1 Relationship of evaluation target and points to consider in setting the definition of Matching with Ground Truth

Relationship of evaluation target (Output of label and the product submitted for registration)	Points to consider in setting the definition of GTs
Face to face	Generally, it is defined using Dice coefficient, IoU, Simpson coefficient, etc. Based on the ROLE of the product submitted for registration, it should be explained that there is a significant degree of overlap of faces, taking into account the importance of correctness of information on ROLE.
Face to point (*)	It is generally defined by the point being included in the face. It is necessary to pay attention to whether the evaluation will be unduly advantageous for the product submitted for registration.
Point (*) to point (*)	It is generally defined by the distance between points. In consideration of the ROLE of the product submitted for registration, it should be explained that it is the distance whose significance can be explained.

* The output specification includes a circle with a certain size, etc.

Similarly, a test to evaluate clinical significance would be expected to assess the degree of agreement between the determination results of the reader and the GTs. Considering the purpose of the test and the method of presentation of the determination results of the reader, the validity of the definition of GTs should be explained with reference to Table 1.

4.6 Endpoints

In general, the variable that provides the most appropriate and convincing evidence directly related to the primary objective of the test is set as the primary endpoint, and the variables that can provide supplementary measurements related to the primary objective or measurements of effects related to secondary objectives are set as the secondary endpoints. Depending on the ROLE of the product submitted for registration, both sensitivity and specificity may be important for some products. In this case, it is necessary to consider setting it as a composite endpoint.

Even for the secondary endpoints, if there are results that may cause doubts about the clinical utility, they may become a major problem in obtaining approval for medical devices. Considering the overall results of the product submitted for registration, it is necessary to examine whether the usefulness and safety in the ROLE are ensured and whether necessary measures such as provision of information are sufficient.

In addition, it is necessary to consider the specifications of the product submitted for registration for what is used as the primary endpoint in the verification test. For example, consider a product that detects multiple findings and classifies and presents the detected findings. By reference to the detected

and classified contents presented, if the final classification result has a high impact on patients (for example, it is assumed to be used in an emergency and used in a position where the treatment of patients is changed depending on the classification result), it is also necessary to carefully consider verifying that each classification result can be detected appropriately.

4.6.1 Criteria to evaluate usefulness

It is necessary to set criteria that can explain the items to be evaluated from the ROLE of the product submitted for registration and the purpose of the test. Examples of criteria used for evaluation of the CAD function are shown below. If the validity of the evaluation of the product submitted for registration can be explained, the criteria other than those shown below may be used.

(1) AUC, FOM

As a criterion used for evaluation of diagnostic support performance, the evaluation by Area under the curve (AUC) using Receiver operating characteristic curve (ROC) is well known. However, when the correctness of position information, which is detected, is also evaluated, it is not possible to evaluate by AUC, and therefore the evaluation is often made by figure of merit (FOM) using free-response receiver operating characteristic curve (FROC).

AUC and FOM are useful when evaluating the diagnostic performance of the two diagnostic methods. For example, it is assumed that the reading results are compared between the physician group without CAD and the physician group with CAD in the reading test. However, when the performance of the product submitted for registration itself is evaluated based only on the AUC and FOM of the product, it is often difficult to explain whether the product has the performance that can be said to be clinically useful. In this case, it is better to explain based on the sensitivity/specificity, etc. shown in (2).

Even when the diagnostic performance of the two diagnostic methods is compared, interpretation may be difficult if the two curves intersect. Although the characteristics of diagnostic performance can be examined from the curve trends, interpretation may be difficult in terms of comparing the relative merits of the two diagnostic methods. It is necessary to examine whether it is possible to evaluate by AUC and FOM based on the results of preliminary tests conducted in advance.

(2) Sensitivity/specificity

Sensitivity/specificity is an indicator of the percentage that each of the positive and negative status is appropriately determined as the label. If the sensitivity/specificity of the conventional diagnostic method is shown in literature, etc., it is a relatively easy indicator to directly explain the clinical utility of the product submitted for registration by comparing with it.

When evaluating by sensitivity/specificity, the meaning of the value calculated differs depending on what was defined as true positive/true negative in the test, what condition was defined as GTs, etc. In addition, the analysis unit (case unit, finding unit, etc.) also changes the meaning of the calculated value. Therefore, it is necessary to clarify the definition of sensitivity/specificity in the test. When comparing with conventional diagnostic methods, etc., it should also be examined whether the

definition of sensitivity/specificity to be compared is consistent with the definition of sensitivity/specificity to be evaluated for the product submitted for registration, and whether it is possible to compare.

(3) Accuracy, positive predictive value/negative predictive value

The accuracy is an indicator that indicates the percentage of patients who were properly assessed as true-positive or true-negative in the entire population evaluated. The positive predictive value/negative predictive value is an indicator showing the percentages of positive/negative judgments of the evaluated samples that can be appropriately judged. It should be noted that these indicators are affected by the positive rate of the population evaluated. When comparing with other diagnostic methods, etc., it should be confirmed whether there is any difference in the positive rate from the comparable methods. In addition, when the performance of the product submitted for registration is specified by these indicators, it is also necessary to specify the performance together with the information on the positive rate of the analysis target. In addition, definitions of these indicators as well as sensitivity/specificity (including analysis unit) need to be clarified.

Table 2 Sensitivity/specificity, accuracy, positive predictive value/negative predictive value

		label	
		Positive	Negative
Analysis results of the product submitted for registration	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

Sensitivity : $\frac{TP}{TP+FN}$

Specificity : $\frac{TN}{FP+TN}$

Accuracy : $\frac{TP+TN}{TP+FP+TN+FN}$

Positive predictive value : $\frac{TP}{TP+FP}$

Negative predictive value : $\frac{TN}{TN+FN}$

4.6.2 Change for the worse and change for the better

In the reading test, the results of diagnosis without CAD are compared with the results of diagnosis using CAD to evaluate the usefulness of using CAD. It can be evaluated whether the overall diagnostic performance is improved by FROC, etc., but in general, false-positive results increase as the detection of true positive increases (sensitivity increases but specificity decreases). It is necessary to discuss the details of cases in which judgment was improved by using CAD (change for the better) and cases in which judgment was worsened (change for the worse), and in particular, for the change for the worse, the necessity of provision of information such as attention calling, etc. should be considered after examining the trend.

4.6.3 Subgroup analysis

Subgroup analysis should be performed as appropriate so that information can be provided to physicians, etc. who use the performance and limitations of the product submitted for registration in order to have them correctly understand and use the product. Examples of perspectives for subgroup analysis are shown below. Depending on the characteristics of the product submitted for registration, selection, addition, or adjustment should be considered as appropriate.

- By imaging modality vendor
- By image processing condition and imaging condition
- By patient background
- By size and properties, etc. of findings
- By the department of readers, etc. in reading tests and by years of experience

4.7 Readers in reading tests

The purpose of the reading test is to evaluate the value of introducing the product submitted for registration after reproducing the actual clinical practice as much as possible. Therefore, it is desirable that participating readers (depending on the product submitted for registration, participants in the reading test are not necessarily physicians) are also selected so that the intended users of the product are represented. For selection, the department, experience/skills, presence or absence of specialists, etc. should be considered.

Depending on the development concept of the product submitted for registration, some products aim to equalize the diagnostic performance by improving the diagnostic performance of physicians who are not familiar with diagnosis. In this case, the main user of the product submitted for registration is a physician not familiar with diagnosis. Such products may not restrict the use of experienced physicians. In this case, the target of the reading test should be a group of readers who can use the product, including experienced physicians. On the other hand, since the diagnostic performance of experienced physicians (although not limited by the experience of physicians) is sufficiently high and diagnostic support is virtually unnecessary, if it is almost unlikely that the product submitted for registration will be used by experienced physicians, it is possible to consider using the evaluation of a limited group of physicians in the reading test as the pivotal. Even in this case, it should be discussed that the use of the product by an experienced physician will not worsen the diagnostic results and appropriate measures should be taken.

4.8 Other

Based on the test results (see mainly Sections 4.6.2 and 4.6.3), the necessity of providing information to ensure that users correctly understand the performance, limitations, etc. of the product submitted for registration should be examined. For example, the following contents are assumed.

- False-positive/false-negative for out-of-target findings
- False-positive/false-negative for parts that overlap with other organs

5. Additional Points to Consider for Products Using Machine Learning

Recently, medical devices using machine learning such as deep learning (hereinafter referred to as "MLMD")) have been actively developed, and the development of CADe, an MLMD, is also becoming mainstream. Discriminators developed using machine learning have features that vary the output nonlinearly with respect to the input (the nature of which output results can vary significantly with respect to changes in input) and have important properties for solving complex pattern recognition problems. On the other hand, because of its characteristics, it is often difficult to predict the behavior of unknown data, and there are characteristics such as that unexpected errors may occur and that over-fitting may reduce the performance. In addition, with regard to machine learning, particularly deep learning, it is difficult to interpret the process of making decisions based on neural networks due to its characteristics. Therefore, it is difficult to explain that the quality of outputs is secured only with principles (such as detection algorithms to be implemented), design specifications, etc. that must be specified as approved matters to secure the performance of usual medical devices.

Therefore, at present, the evaluation is focused on confirming whether appropriate outputs are obtained for input based on ROLE, instead of detailed examination of the contents of the established network. That is, its appropriateness as a "medical device" is evaluated, regardless of the presence or absence of development using machine learning. Based on the above, in order to demonstrate the efficacy and safety of the product submitted for registration, it is required to verify the performance, etc. of the product submitted for registration in consideration of a medically and statistically valid method as described above.

On that basis, additional considerations to CADe developed using machine learning are summarized.

5.1 Points to consider for test data set

Many of the products submitted for registration were developed on the premise that they will be used at any institutions in Japan and approval will be obtained. Therefore, it is necessary to evaluate whether the quality, efficacy, and safety of the product submitted for registration are ensured at any institutions in Japan. In other words, the generalizability of the test results will be a review issue.

Regarding the evaluation of the CAD function for the product submitted for registration which is an MLMD, considerations for the bias possessed by the CAD function and considerations for the test variations to be considered because the design is inductive are described below.

5.1.1 Relationship with training data

The test results of the analysis results of the CAD function of the product submitted for registration which is an MLMD are affected by the bias of the test dataset and the bias of the model. The processing of bias possessed by test data should be considered as evaluation of usual medical devices regardless of whether or not the product for registration is an MLMD (Generally, it is controlled by data collection facilities and inclusion/exclusion criteria).

On the other hand, in the case of an MLMD, the bias of the model should also be considered.

For example, if the development data and test data are collected at the same institution, the test results are strongly influenced by the bias of the institution such as the patients who visit the institution,

imaging model, imaging method, and medical treatment policy/system. Therefore, there is a possibility of divergence from the results obtained when similar tests were performed with data from other facilities. This may make it difficult to discuss the generalizability of the test results.

Another example is that if the method of annotating labels in development data and test data contain many subjective elements of the creating physicians, it is approved as an evaluation for imitating their judgment, but it may be the evaluation that does not allow general physicians to adequately detect what they should detect. This may also make it difficult to discuss the generalizability of the test results.

As described above, even if CAD has the same clinical ROLE, the judgment on the appropriateness as test data may differ depending on the relationship with the development data. Therefore, it is necessary to explain the relationship between development data and test data in an organized manner with reference to Table 3, etc.

Table 3　Example of organization of relationship between development data and evaluation data

	Training data	Test data
Data collection location	○○ University Hospital ○○ Hospital	
Patient background	Patients undergoing medical checkup/patients requiring examination in the primary screening	
Imaging modality, Imaging conditions		
How to establish supervised data/labels	Positive: ○ specialists in ○○ in Japan read images only using image information, and the location considered ○ was defined as ○. *Negative:* • • •	
Relationship with training data		Data sets containing part of training data/another data set collected in the same institution obtaining training data /another data set collected in a different institution than training data, etc.

Since it is difficult to quantitatively discuss the influence of bias, it is desirable to collect and evaluate test data sets that are considered unlikely to be affected by bias. For example, the following items should be organized.

- Institutions to collect training data and test data should not be duplicated.
- Creators of supervised data and the label should not be duplicated.

If it is difficult to take any of these measures, the generalizability of the test results should be explained by examining the presence or absence of potential bias caused by the measures and their influence.

5.1.2. Consideration for variation

If the product submitted for registration is functionally designed a priori by the developer, it is relatively easy to identify the factors that do not affect the product (or to explain the validity), and thus factors that do not need to be included in the test data can also be explained relatively easily. On the other hand, in the case of an MLMD, it is difficult to identify the factors that do not affect the product submitted for registration (or to explain the validity) because the function is designed inductively. Therefore, it is necessary to assume various factors that can be entered in the product submitted for registration and include them in the variation of test data set. Examples of factors are described below.

(1) Examples of clinical factors

- Severity of target disease
- Type of detection target
- Number of detection targets
- Anatomical location of detection target
- Patient age, sex

(2) Examples of nonclinical factors

- Imaging modality
- Imaging conditions
- Imaging parameters

Due to the diversity of the training data set at the time of development, factors that do not affect the product submitted for registration may be discussed (for example, consideration that there is no influence of differences between models because training is performed with data obtained from various models). However, it is not possible to explain that the stability of the analytical performance for these factors is secured only with the explanation included in the training data in general. In principle, please consider a plan to collect test data so that various variations are included by reference to the above examples of factors.

End of document