PSEHB/MDED (Yakuseikisinn) Notification 0523 No.2
May 23, 2019
Appendix 4

Guidance for evaluation of artificial intelligence–assisted medical imaging systems for clinical diagnosis

1. Introduction

Computer-aided diagnosis (CAD) is used as a "second opinion" to help doctors make clinical diagnosis decisions by indicating the results of quantitative computer analyses of medical images, including radiographic images.

Guidance on medical devices for CAD has been published in the third attachment of the "Publication of guidance for medical devices with emerging technologies" (Notification No. 1207-1 by the Director of Office of Medical Device Evaluation (OMDE), Evaluation and Licensing Division (ELD), Pharmaceutical and Food Safety Bureau (PFSB), Japanese Ministry of Health, Labor, and Welfare (MHLW), dated December 7, 2011).

However, it is difficult to apply the guidance to such devices with current artificial intelligence (AI) technologies. The performance of these devices could be improved after they are marketed with additional data (post-market training), using machine learning (ML)-based technologies, which are facilitated by recent significant advances in the field of computer science. In addition, it has been reported that new, unprecedented challenges have surfaced during the evaluation and operation of devices with AI related to deep learning. Although deep learning is one of the ML-based technologies for AI, it has recently gained significant attention because the algorithm for calculating the final output has a "black box" nature and is not clarified. Because of its "black box" nature, its performance, especially its modified performance after post-market training, can only be evaluated by adequate verification of the outputs. Therefore, new points of view are required to evaluate such AI-assisted medical devices for CAD. The recommended approaches cover how to evaluate the performance of such devices before and after marketing, and how to ensure high learning quality. Factors, such as the source or type of the data, authenticity and bias in the learning data, and the evaluation method for the post-market performance changes should be considered.

Given these circumstances, this guidance summarizes the issues and points to evaluating the efficacy and safety of the medical imaging system for CAD utilizing AI technology for its market approval, while considering the above-mentioned facts.

2. Definition and explanation of terms

The terms used in this guidance are defined as follows.

(1) Medical imaging system for clinical diagnosis

A system or software with a computer-aided detection (CADe) or computer-aided diagnostic (CADx) function that provides information that is used as a reference in clinical practices when doctors perform image-based diagnosis (interpretation) using various medical imaging devices.

Note The type of assistance provided for the user (for instance, the doctor who uses the system for assisting diagnosis in clinical practice) is classified as "first reader," "second reader," or "concurrent reader."

1) CADe

A stand-alone software, or a device incorporating the software, with functions that automatically extract suspected lesions on images and mark their locations by analyzing the medical image data, or both the medical image data and results of other medical testing, in order to assist in the detection of lesions and/or abnormal values.

2) CADx

A stand-alone software, or a device incorporating the software, that outputs quantitative data as a numerical value or graph, such as the differentiation between benign and malignant lesions and a prognosis based on the stage of lesions.

Note 1 A software or a device incorporating the software with functions that support diagnosis by providing candidates of diagnosis and/or results of risk assessment is also included in this category.

Note 2 The definitions of CADe and CADx are cited with partial modifications from the aforementioned OMED/ELD/PFSB/MHLW notification.

(2) Artificial intelligence (AI)

A computer system or software that mimics inference or learning by advanced human intelligence.

Note The guidance presented in this study is directed toward those systems with functions that change the system itself after the clinical use has started (post-market training function) with additional data through machine learning (for example, deep learning).

(3) Machine learning (ML)

A technology that has been studied significantly as a part of AI technologies and is a method that enables a computer (software) to mimic human learning functions, including various algorithms, such as the discriminant analysis, decision trees, neural networks, and support-vector machines.

Note Typical applications include classification and regression; since these are consistent with the purpose of CAD, they have been utilized for developing various CAD systems.

(4) Deep learning

A part of the broader family of ML that has recently gained a considerable amount of attention and is characterized by the use of a large volume of data that is set up as a deep-layered neural network, which includes deterministic models, such as hierarchical networks, and stochastic models, such as Boltzmann machines. Its algorithms for training the networks are categorized as "supervised learning," "unsupervised learning," or "semi-supervised learning," and they can include reinforcement learning.

Note Since the publication of research by Hinton et al. in 2006, deep learning has been used in software programs such as "Alpha Go", which defeated the world champion of "Go", and in other software programs that have achieved higher rankings in various competitions in the field of computer technology. Because of these outstanding results, deep learning has been studied thoroughly in the field of computer science. One of the reasons for the significant improvement in its performance is the appearance of a new technology enabling the training of deep-layered neural networks. Another reason is the progress of big data management technologies and parallelly distributed computing technologies.

(5) Transfer learning

A part of the training technologies of ML proposed more than 20 years ago which is defined as "a technology to effectively obtain an effective hypothesized model for the new task by applying the existing model learned from one or more different tasks."

Note Transfer learning has been actively used in deep learning in recent years. Some of the specific examples for its usage are as follows:

1) Using the neural network that was developed by a large-scale dataset as a feature extractor, transferring the resulting features from new data into tasks that are completely different from its original purpose. For example, inputting medical images to a neural network, which has already been developed by learning with general natural images, and utilizing the outputs from an intermediate-layer of the network as a feature amount to assist the diagnosis from those medical images.

2) Preparing a trained neural network with another purpose, and relearning by using the training data of a different task. The possible examples include a case where the purpose is recognition followed by diagnosis using CT images, where the CT images and annotation data are applied for re-training the neural network trained with the natural images, as explained in example 1). In such a case, a neural network with better performance is obtained with fewer CT images and lesser annotation data in comparison with a neural network that has to be trained from scratch.

3. Scope of the guidance

This guidance describes principles to evaluate the quality, safety, and efficacy of medical imaging systems assisted by current AI technologies for clinical diagnosis. The purpose of the systems is to assist medical doctors in making decision for clinical diagnosis with the medical images obtained from various devices. The functions used in the systems are classified as follows (CADe or CADx are hereinafter referred to as "assistance systems"):

- Functions that detect only the sites with suspected lesions without specifying the name of the disease (CADe).
- Functions that present the names of the possible diseases in addition to the sites with suspected lesions (CADx).
- Functions that present the names of the possible diseases and their severity (clinical grade).

This guidance is applied to the assistance systems whose performance could be changed intentionally by applying post-market training of ML-based technologies, irrespective of the type of algorithm or application form of AI technologies. It is worth noting that the assistance systems addressed in this guidance are for assisting doctors in making diagnoses; thus, the basic requirement is that any unexpected behavior or malfunction by the systems shall be detected by the user or notified to the users.

However, it is extremely difficult for the manufacturers or distributors to control the quality of the assistance systems with AI when the changes in performance have been caused by post-market training. This is because there is the possibility that changes in performance may vary between facilities because of post-market training by the end user or changes may occur automatically. The sheer volume of problems that need to be resolved renders it difficult to include assistance systems as a subject of present guidance. Commercialization of these systems has been realized owing to advances in AI technology, and further development in the assistance systems with AI is highly expected. Therefore, a separate attachment summarizes the basic approach for ensuring the quality, safety, and efficacy of the devices.

If it is difficult to ascertain whether or not the guidance is applicable to any assistance system, it is recommended to consult the Medical Device Evaluation Division (MDED) Pharmaceutical Safety and Environmental Health Bureau (PSEHB), MHLW.

4. Role of the guidance

The recent advances in the machine learning AI technology, especially deep learning,

which enables automatic training by the existing data, has allowed us to accept the clinical use of assistance systems with AI. AI can perform repeated training during a shorter period of time because of the development of frameworks that continuously collect large amounts of data and significant improvement of computation performance. It is now technically feasible for the assistance systems to automatically collect clinical data and to improve their performance by post-market training. This leads to the expectation that implementing assistance systems will improve their performance either in a stepwise or in continuous manner after marketing. In contrast, the automated training systems can possibly degrade performance by inputting improper data. Therefore, there are arguments both for and against these possible functions

This guidance only describes the points to be considered, including the problems that need to be solved for market approval of these assistance systems. Thus, this guidance is an informative document for the market approval applications. In addition, the revision of this guidance will be required in the future based on future technological innovations or accumulation of knowledge.

Assistance systems should be evaluated flexibly based on a scientific rationale and with a full understanding of the individual methods for their construction, intended use, and characteristics. In addition to this guidance, the above-mentioned “Guidance for computer-aided diagnosis devices” and related guidelines (both domestic and foreign) should be good references for their evaluation. For instance, refer to the office memorandum titled “Guidance on applying for a market approval of medical device programs” provided as an attachment to the “Publication of guidance on applying for a market approval of medical device programs” circulated by the Medical Device and Regenerative Medicine Product Evaluation Division (MDRMPED), PSEHB, MHLW, dated March 31, 2016 for more information.

Note Factors for post-market training by deep learning include the weights of network connections (coupling), as well as network hyperparameters (such as the number of layers, number of feature maps, convolution filter size, number of units, etc.). One example of the training process is to select an appropriate combination of hyperparameters (including the above) that exhibit the fewest number of errors from the results of exhaustive learning of their various combinations to minimize errors (losses) between training data and output from networks. However, it is recommended to consult with the Pharmaceuticals and Medical Devices Agency if assistance systems with the modified hyperparameters by post-market training for improving their performance are accepted as they may require minor modification or a process for a partial-modification approval.

5. Open problems and direction of their solutions

(1) Black box

When utilizing machine learning (deep learning in particular) for AI, a final process for decision making by a neural network always becomes a "black box" because of its design principles, wherein a clear justification for an output from the AI is sometimes difficult. Therefore, it is difficult to elucidate an "operating principle" to ensure performance with conventional medical devices for their market approval. During the approval process, rather than describing the details of the diagnostic algorithms, the focus should be on the performance evaluation by confirming whether the input yields the required output, which is similar to the evaluation of other medical device programs.

As described above, it is difficult to describe the operating principles and algorithms in detail; thus, it is also difficult to predict the output other than where the performance has been confirmed (i.e., rare cases not included in the training data during development of AI). Considering these risks, to ensure the performance of assistance systems utilizing ML (particularly for those utilizing deep learning), manufacturers should guarantee the performance by indicating that the systems always meet the performance specifications, which have been stipulated by the manufacturers based on scientific and clinical rationales in accordance with the methods and intended use. These values for the target disease should include the detection rate, false positive rate, false negative rate, time required for detection, and other necessary factors (refer to Section 6-3).

Functions that inform any unexpected outputs of the assistance system to the users should also be required. The manufacturers should prepare the functions that easily facilitates the verification of its performance by the users (refer to Section 6-4). For instance, only allowing users who satisfy requirements to be able to determine any unexpected outputs of the system or to provide immediate feedback of usage monitoring under clinical circumstances by which the manufacturers can verify the performance.

It is extremely difficult to completely eliminate unexpected behaviors or misjudgments; thus, operational processes should be considered beforehand, which collect the above information to perform appropriate countermeasures.

(2) Changes in performance

The performance of some assistance systems utilizing AI are expected to improve by altering the diagnostic algorithms from post-market training. However, this alteration may degrade their performance by post-market training because of the unexpected changes in the algorithms. Taking these benefits and risks into consideration, the

assistance systems are required to ensure the quality of the training by defining the requirements for the training data and processes, as well as by continuously verifying their performance.

1) Continuous verification of performance

If assistance systems can change their performance continuously or frequently by post-market training during their clinical use, the assistance systems should be validated each time their performance changes to ensure their quality, safety, and efficacy. During this process, the verification and management of the altered performance is especially important, in addition to the management of the entire process. Therefore, the change in performance levels that may be caused by post-market training should be stipulated in advance in accordance with the purpose within an acceptable range based on clinical and statistical validity. Furthermore, verification methods that can be applied to the devices and measures if the change in performance is beyond the stipulated range should be defined (refer to Sections 6-3 and 6-4). Regarding the data used for verification, the reason and validity for the use should be indicated with a clarification of their source (refer to Section 6-2-2).

2) Quality assurance associated with performance changes

Improving the performance of assistance systems utilizing AI technology requires appropriate post-market training; thus, the training algorithm and data should be clarified (refer to Section 6-2-2).

If an assistant system has an algorithm to learn new data collected via their clinical usages, their performance will be changed just after finishing the training process. In this case, the training process and the performance verifications for the changes in its performance should be implemented by the manufacturers to ensure the quality, indicating a necessity to satisfy the requirements which are similar to those for its original market approval (refer to Section 6-3-1).

Meanwhile, for devices that can change their performance continuously or frequently by post-market training at the same time of their clinical use, their performance may deteriorate contrary to the manufacturer's intention, i.e., it may fall below levels that are clinically unacceptable. To prevent such deteriorations, the assistance systems should be connected to the manufacturer via the internet (refer to Section 6-3-2) so that the manufacturer can perform its adequate verification or measures in accordance with the purpose and risks associated with the assistance systems.

In principle, the manufacturers should ensure the quality of the assistance system even after its performance has been changed. In addition, any additional measures

necessary for ensuring the quality that depend on the risks associated with utilization of the assistance systems should be provided. Some examples include applying measures to the user's computer if the performance deteriorates below the lower limit, conducting a training session for users to prevent misuse or off-label use, and disabling post-market training functions by the users (refer to Section 6-4).

3) Principles on market approval process

Generally, additional processes for market approval may be required if improvement of the performance of assistance systems affects the clinical efficacy as well as its quality or safety; however, the necessity of the process should be determined in accordance with the magnitude of the changes in performance as well as the considerable risks associated with those changes.

An upgrade that improves the performance may produce a new product, or applying such an upgrade may incur costs; thus, it is conceivable that this type of change in performance should not be applied to all relevant assistance systems already sold on the market, thereby resulting in existence different versions on the market. Follow-up systems pertaining to upgrades and a suitable process for market approval of the upgrade should be discussed with consideration of the unique distribution of assistance systems utilizing AI.

(3) Assigning responsibility

For medical devices utilizing AI, concerns about the ambiguity of locating responsibility of their usage have been raised. For assistance systems, there is a risk that doctors entrust the final diagnosis to these medical devices, even though their purpose is only to "aid in the diagnosis." As with conventional medical devices, the manufacturers are responsible for the maintenance and the troubleshooting of assistance systems in the event of the design- or specification-related problems, failures, or malfunctions; however, in order to clarify the responsibility of their users as well as their proper use, it is necessary to indicate the intended use and use method for assistance systems, and to provide measures such as an operation training for ensuring their proper use by users. (refer to Section 6-4). It is worth noting that the notification No. 1219-1 by the Director of Medical Profession Division, Health Policy Bureau, MHLW entitled "Relationship between the use of software programs utilizing artificial intelligence (AI) for assisting diagnosis, treatment, etc., and the stipulations of the Medical Practitioners' Act Article 17", dated December 19, 2018, point out that medical doctors play the major role in diagnosis and treatment even if they utilize such assistance software programs and that they are the ones who are responsible for the

final decision related to diagnosis and treatment.

6. Points to consider in evaluation

(1) Principles

Since the assistance systems to which this guidance applies are different from conventional medical devices, it is difficult to ensure their performance by principles of the mechanism (the implemented detection/diagnostic algorithms) or design specifications. Therefore, it is necessary to define factors that affect their performance, including conditions with which their performance and limitations can be evaluated. Satisfying requirements for their market approval, the performance of assistance systems should be verified utilizing the factors defined in accordance with medically and statistically valid methods. The quality, safety, and efficacy of assistance systems should be evaluated utilizing performance levels set according to their intended purposes with reference to the aforementioned notification entitled "Guidance regarding computer-aided diagnosis devices."

Considering the above-mentioned contents, specific points to consider in evaluation of the assistance systems are described based on state-of-the-art sciences.

(2) General points on principles of detection/diagnosis, learning, and information security of the assistance systems

1) Principles of detection/diagnosis (algorithms)

For the systems utilizing the detection/diagnostic algorithms whose processes can be specified, the final algorithms and an overview of the software programs for the system should be provided for their market approval process. However, for the systems utilizing the algorithms, such as deep learning, whose processes cannot be specified (become a "black box"), it is difficult to provide final algorithms for the detection/diagnosis for the market approval process. In this case, an original network structure of deep learning for the detection/diagnosis and an overview of the software programs should be provided for the approval process. If an advance in technology makes it possible to clarify the final behaviors of a "black box" network, such information should be provided to the extent possible.

2) Training

The assistance systems should have the performance required to achieve their intended purpose by training AI using appropriate training data. Thus, it is necessary to explicitly indicate the content for items that is required for their market approval process with reference to the following examples in accordance with the mechanisms

and defined performances of the assistance systems, as well as to indicate the reasons and validity for using them.

- Learning algorithms and overview of software programs
  (clarify if the learning algorithm is supervised, semi-supervised, reinforcement learning, self-learning, etc.)
- Data (specify the necessary items with reference to the following for training data, validation data[Note 1], and test data[Note 2])

Note 1 Validation data: Data used for determining the hyperparameters of machine-learning algorithms (such as parameters for determining the objective functions of support-vector machines, number of layers of deep learning networks, number of feature maps, convolution filter size, and number of iterations of learning.)

Note 2 Test data: Data used for evaluating/verifying the performance of systems

- Data source (includes clarification of the methods of data acquisition and management. In particular, when using the images and other information as post-market training data, which have been acquired from the use of assistance systems after their marketing, methods to obtain informed consent from patients should be also included.)
- Imaging parameters during image data acquisition.
- Type of clinical data linked with images[Note 3] (includes annotations, such as locations of the lesions, size (including label images), and differentiation results between benign and malignant.)
- Data other than that from clinical images if used [Note4] (image data created by a computer simulation for data augmentation or using a phantom.)
- The number, size, density levels, and other necessary factors of the image data.
- Applicable processing methods if processing (e.g., anonymization) was performed before training.
- Doctors or experts who made the final decision regarding clinical data.
- Those responsible for the final decision if data other than clinical data were used.
- Schemes for managing the test data by completely severing its ties from the learning process.

Note 3 Hitherto, almost all assistance systems have been developed utilizing medical images which have annotations with high clinical reliability. However, advances in technology in recent years have enabled the use of natural or artificial images, as well as linked data with low reliability in terms of

clinical information (weak label data), as training data for developing the assistance systems. However, if such data are to be used as training data of the systems, its validity and justification for use should be provided. Regarding points to consider for the test data, refer to Section 6-3 in accordance with the evaluation stage.

Note 4 When utilizing transfer learning for the assistance systems, the use of natural images, medical images obtained by different imaging modalities, or images constructed artificially as training data is expected. Detailed information should be provided regarding the target legions, types, and number of images if natural images are used. If medical images obtained by the different modalities are used, similar information to those obtained by a modality for the final system should be provided. If artificial images are used, methods of their construction and other necessary information should be provided in detail.

If the images used as training data for the assistance systems are acquired from a database, the use of the database should be validated. The following are examples of items related to the database of which a description may be required for market approval application of the systems. It should be noted that not all but necessary items should be appropriately selected or added in accordance with features of the used database:

- Overview of database administrator (academic society, certified anonymous processed medical information creator, etc.), organizational structures, etc.
- Business plans documented by the administrator.
- Business content outsourced to a third party by the administrator.
- Types of data stored in the database.
- Outline of database and design.
- Procedures involving the management of data and their operational status

and so on.

All items related to the above-mentioned training should be organized into common and different requirements and separately described between training during pre-market development and post-market development of the system. For example, if the machine-learning algorithms for training during the design and development is different from those for post-market training, each of them should be clarified and separately described. If the hyperparameters of the machine-learning algorithm are re-determined by post-market training, the possible ranges of their changes should be defined in advance, and the changes of those hyperparameters after the post-market training should be validated to be within the defined ranges. Furthermore, changes of the neural network after post-market training should be recorded in each change of

the performance for necessary validation of the systems in the future.

3) Use environment and information security

When using the AI technology, the following items should be described after clarifying the type of AI, either on-premise or cloud.

- Methods of data transfer from imaging modalities.
- Possibility of interference with imaging modalities or other types of software for the modalities by installing a necessary software for use of the systems.
- If an assistance system is connected to an external device, such as the Internet, measures required for ensuring its cyber security are prepared with referring to the joint notification No. 0428-1 by the Director of MDRMPED, Minister Secretariat, MHLW and by the Director of Safety Division, PFSB, MHLW entitled "Ensuring cyber security in medical devices", dated April 28, 2015.
- If the clinical information that was obtained through the usage of the assistance system is converted into post-market training data, proper management and protection of personal information are placed with referring to the following guidelines:
  - "Guidance for the appropriate handling of personal information in medical and nursing-related providers," dated April 14, 2017, Personal Information Protection Commission and MHLW
  - "Guideline for the safety management in medical information systems, 5th ed.," May 2017, MHLW

Note If necessary, refer to the Act on the Protection of Personal Information as well as the Act Regarding Anonymized Medical Data to Contribute to R&D in the Medical Field.

4) Conditions of imaging modalities for assistance systems

For assistance systems to fulfill their predetermined functions, a format, a resolution, and other parameters related to the medical image data obtained by the modalities of the systems should satisfy the necessary requirements of the systems. Since the performance of the assistance systems may be influenced by the capturing methods and the parameters of the medical image data for the learning process of the AI, it is necessary to define them even if the system can be utilized with all specific imaging modalities on the market.

(3) Evaluation of safety, quality, and efficacy

1) When applying for market approval

The evaluation of an assistance system should be performed in correspondence to its purpose and risks; thus, it is necessary to determine the standard values (detection

rate, etc.) for ensuring its quality, safety, and efficacy with appropriate validity, based on its purpose and risks, before performing its evaluation.

Other than those items listed in Section 6-2, issues specific to assistance systems that require clarification in various evaluations include the following:

- Functions of connected imaging modalities (especially for a system that can utilize an existing image modality).
- The performance verification protocol of the systems, including those responsible for the verification.
  - If the performance verification is performed as a clinical trial, protocols to determine the diagnosis results of its subjects should be validated to ensure accuracy of the trial.
  - If a retrospective verification is performed, test data of the verification should be validated with necessary information.
- The type, source, and validity of the test data used for the performance verification (management methods and other information should also be clarified if necessary).
- Defined ranges of performance changes that may occur after marketing (lower limit for the detection rate, limits for the false positive rate, false negative rate, etc.) and clinical or statistical data ensuring the validity of these definitions of the ranges.
- Measures for ensuring that the efficacy and safety are secured within the defined ranges even when performance changes have occurred after marketing (methods to obtain training data, methods for verifying the performance after the change, etc.).
- Problems that may occur with performance changes,

  etc.

Sets of test data used for performance verification should be independent of those of training or validation data for the systems. Moreover, when considering the characteristics of the subject group, the number and quality of the test data should be significant enough to be able to explain their validity for the verification. However, it is difficult to define the requisite number of test data uniformly because the necessary number of test data for assistance systems varies depending on the method and subjects of the machine learning for the systems as well as their purpose. Therefore, not only the information regarding the method of performance verification of the assistance systems and test data utilized, but also the schemes for validating

performance that include the test data should be documented.

2) After post-market training

Regarding the performance changes that occur in the assistance systems after marketing, the performance change by post-market training data should be validated after confirming that it satisfies the predefined ranges from a clinical or statistical perspective by the performance verification methods specified when applying for their market approval.

When assistance systems are utilized within the existing imaging modalities, the quality of the image data obtained may vary depending on the performance of the imaging modalities. Therefore, requirements of image data for post-market training should be defined so that measures can exclude data from the post-market training that does not meet the requirements. Requirements of the performance verification process and sets of test data for the systems after post-market training are essentially the same as those used for the first application for market approval mentioned in the previous section. Furthermore, it is also required to clarify which medical doctors or specialists are responsible for the determined diagnosis of data for the post-market training as well as for test data to evaluate the performance of the systems with post-market training, as described in the previous section.

For assistance systems that are connected to AI through a network, a mechanism informing their users of any changes that have occurred in their performance would be preferable so that the users can obtain a complete understanding of the details regarding their required performance and operation.

The necessary considerations for each method of post-market training for the performance changes of assistance systems are shown below.

1: Stepwise changes in the performance with each upgrade of the systems

This applies to the assistance systems whose performance will be changed by a post-market training with data obtained from clinical practice. In this case, collection of the data as well as the post-market training of the systems should be performed by manufacturers in a suitable manner (e.g., minimizing data biases and collecting enough data based on clinical/statistical validity). If data other than those obtained from clinical practice are used, such use should be clarified and validated when applying for market approval of the systems.

In such cases, the manufacturers are responsible for the evaluation of the revised performance, in addition to ensuring the safety and quality associated with such changes in each upgrade. For the evaluation, the aforementioned “Guidance for

computer-aided diagnosis devices" should be referred to as when applying for market approval of the systems.

Furthermore, the procedures for market approval required for the assistance systems, for which performance can change by post-market training, may vary depending on the principles and characteristics of the systems; thus, the anticipated ranges of the performance change should be determined by consultation with the Pharmaceuticals and Medical Devices Agency prior to applying for market approval.

2: Continuous or frequent changes in the performance by post-market training associated with clinical practices

This applies to the assistance systems that enable post-market training of data obtained by their clinical practices continuously or frequently, thereby causing continuous or frequent changes in their performance. In this case, the AI for the systems should be managed by the manufacturers through a network for the respective assistance systems. In such cases, it is especially important to ensure that the changed performance and quality of the assistance systems by the automatic post-market training are within the predefined ranges at the time of market approval. Therefore, the training processes and methods for managing them should be precisely documented. In addition, mechanisms for keeping the changed performance within the predefined ranges should be set up in advance, and the performance should be periodically verified by the manufacturers. Furthermore, risk management for possible problems occurring with such upgrades, such as unexpected behaviors or an incorrect diagnosis, are necessary to prepare their solutions.

(4) Risk management

1) Principles

The risks associated with assistance systems include "wrong information" presented by the systems which cause a deviation in performance from the predefined ranges as a result of the post-market training. For assistance systems wherein a stepwise change in the performance is intended for each upgrade, the manufacturers should market them only after the problems caused by the performance changes are solved so that the risks can be minimized. In contrast, for assistance systems wherein a continuous or frequent change in the performance by immediate post-market training is intended, there is a high risk of "wrong information presentation" due to the performance changes; thus, a specific mechanism for avoiding these risks should be implemented.

In addition to the above-mentioned risks, off-label use can increase the potential for

unexpected behaviors of the assistance systems which are insensible for the users; thus, measures for preventing off-label use should be provided.

2) Measures for the risks

To reduce the above-mentioned risks, assistance systems that require measures should be equipped with the following functions that either ensure performance based on which ranges are defined during the application for market approval, or specify the steps for the measures depending on the purpose of the assistance:

- A function for saving a detailed log (use environment, diagnostic result, persons responsible, etc.) for images used as post-market training data (refer to Section 6-2).
- A function for periodically assessing the performance by self-testing using the internal data of the assistance system, where the performance at that time is shown to the users, to confirm a minimum performance standard.
- A function for restoring the AI to the initial state at the time of purchase or to the state immediately prior to the change when the above functions have confirmed that the above-mentioned risks have occurred.
- A function for terminating the assistance systems when necessary when the above-mentioned risks occur.
- A function for providing sequential feedback of the circumstances during clinical practices, thereby enabling the manufacturers/distributors to verify the performance as appropriate.

The manufacturers should limit the users to healthcare professionals, such as doctors with expertise who correspond to the intended purpose of their assistance systems and receive education (a training session) on how to use them. Furthermore, it should be clarified to their users that only medical doctors can make the final decisions about definitive diagnoses and other medical practices. Appropriate measures, such as a user training session, should also be provided to prevent misuse by users or off-label use. The measures such as user training would preferably be conducted in accordance with the relevant academics.

7. Examples of additional points to consider according to the characteristics of the equipment

(1) For assistance systems where changes in their performance by post-market training causes a higher risk in some assisted diagnoses than in others

For example, for a pathological diagnosis, the diagnosis made is the final decision in

clinical practice. Therefore, regardless of the type of algorithm, changes in the performance after marketing of the assistance system may decrease the detection rate and/ or accuracy rate, thereby entailing a high clinical risk. Since the quality, safety, and performance of such assistance systems need to be sufficiently ensured, pathologists should be involved in every process of the post-market training, verification of the performance changes, and validation of the systems.

(2) For assistance systems applying to video images

For assistance systems with video-imaging modalities such as gastrointestinal endoscopic devices and ultrasonic diagnostic imaging, details of how to use the video images as the training data for AI should be clarified. Accordingly, in addition to stating the format of the video and how it is acquired, the process of converting the data into training data should also be clarified.

Moreover, capturing the conditions and quality of the video images of such modalities depends on the skill of the users; thus, the conditions that allow the images to be used as training data for such assistance systems should be defined and their validity should be clarified.

-

Draft principles for artificial intelligence-assisted medical imaging systems for clinical diagnosis where users can perform post-market training to change their performance

1. Introduction

The main text of this guidance for medical imaging systems for clinical diagnosis assisted by current artificial intelligence (AI) technologies (assistance systems) was discussed and developed by the working group, which determined if the guidance draft should include all the various feasible assistance systems obtained through the remarkable development of modern computer technology in recent years. In particular, the working group raised a possibility that recent developments in computer technology enable to develop assistance systems with on-premise type AI in a general purpose PC with the same specifications as a current "supercomputer" whose performance can be changed by post-market training of deep learning algorithms (which has gained considerable attention in recent years) by users belonging to clinical facilities. Therefore, discussions have been carried out to determine if the scope of this guidance should include such assistance systems.

Such cases include assistance systems wherein it is difficult for the manufacturers to manage quality after a change in their performance occurs due to post-market training, systems whose performance can be changed by each facility's or each user's post-market training, or systems capable of learning with unsupervised data. As a result of the discussion, these assistance systems have been excluded from this guidance because they involve many problems that should be solved to manage their quality, safety, and efficacy. However, as aforementioned, since it is possible that the remarkable developments in AI technology enable the development of these assistance systems in the near future, this annex summarizes the draft principles for ensuring their quality, safety, and efficacy.

2. Problems and solutions

(1) Performance changes

For assistance systems whose performance can be changed by post-market training by the users who do not fully understand the concept, design, mechanisms, etc., their performance could change significantly from what the manufacturers developed and intended, resulting in a deterioration of their performance to levels lower than those clinically acceptable. To prevent such deterioration, verifications of the performance

using test data with a suitable quality and amount, as well as providing measures that consider the risks of such assistance systems, are required. Some of the examples of those verifications are shown below.

– Requirements to ensure thorough quality assurance by the manufacturers.
– Implementing the same quality management system as the manufacturers in the clinical facilities.
– Conducting training sessions for the users.
– Providing additional measures according to the probable risk if the performance deteriorates below the lower limit.

Because the purpose of the assistance systems is to "assist diagnosis," the above-mentioned risks caused by the deterioration of the performance can be avoided by a doctor making the final decision for the diagnosis. However, implementing a manufacturer function or procedure that addresses all the risks (as necessary) that may have occurred would be required.

(2) Assigning responsibility

In principle, the doctors should be responsible for medical practices, off-label use, and reporting problems associated with the use of such assistance systems. However, the set of responsibilities for the manufacturers and doctors should be clarified, respectively, as it may be impossible for some systems to apply this principle uniformly.

3. Points to consider for the evaluation (evaluation of quality, safety, and efficacy)

The following issues should be considered for systems where the users can perform post-market training at each facility, and for systems that enable automatic post-market training with unsupervised data (self-learning).

(1) For assistance systems which enable the post-market training at each facility

This applies to the systems that enable the users to perform post-market training of data obtained by clinical practices (including diagnosis utilizing the system) at a specific medical facility. In such cases, possible biases in the post-market training data may cause the performance to change in various ways depending on the characteristics of the patients of each medical facility, thereby rendering it difficult for manufacturers to manage the quality of the systems. In addition, when the users implement the post-market training of the system, the performance could change in a direction that differs significantly from what was designed and intended by the manufacturers. Therefore, in addition to the points described in the main text of this guidance, the manufacturers should ask clinical facilities to prepare the same quality management system to ensure

appropriateness of the performance changes by post-market training as required for themselves, with consideration of the following points:

- Details on post-market training protocols, especially those for collecting post-market training data, which is implemented for the post-market training by the users at each facility, as well as to ensure their quality control.
- Details of protocols for verifying the performance of the system that has changed after post-market training. Not only the results of the performance evaluation of the test data for market approval, but also the details regarding the collection methods, protocols, and results of the quality and performance evaluation of the system should be clarified when the new test data is added for the verification.

In addition to post-market training, users should ensure that the performance changes of the assistance systems are within the predefined ranges and their safety and quality associated with those changes is also acceptable in accordance with those accepted at the time of their market approval. Therefore, relevant evaluations should be equivalent to those at the time of implementation by the manufacturers. For such evaluations, clinical facilities purchasing the assistance systems should prepare a quality management system by clarifying a responsible individual within the facility who satisfies the following requirements and documenting responsibilities of the individual, procedures of performance evaluation and quality control and other necessary requirements.

- A responsible individual should be familiar with the principles and training methods of the relevant assistance systems.
- A responsible individual should be able to validate the post-market training data.

As described above, the ranges of performance changes should be within those predefined by the manufacturers for market approval of the assistance systems; thus, the performance should be evaluated by the verification methods defined by the manufacturers for their market approval. In addition to the matters listed in Guidance 6-4 “Risk management,” it is required that the assistance systems are equipped with mechanisms to prevent changes in performance from going beyond the ranges predefined by the manufacturers, e.g., a function that can immediately inform manufacturers of the occurrence of these risks.

(2) For assistance systems that are capable of automatic post-market training with unsupervised data (self-learning)

In October 2017, it was reported in Nature that Google’s *Go*-playing AI could conduct

continuous performance changes by automatic post-market training with unsupervised data (self-learning). Therefore, it is possible that the assistance systems, of which performance can be changed by automatically "self-learning" in parallel to their clinical practices without support from medical doctors and the manufacturers, will be developed utilizing a similar AI in the future. In addition, it is anticipated that assistance systems utilizing such on-premise AI will be developed and their performance changes can be verified at each medical facility without any support from the manufacturers. However, state-of-the-art technology cannot clarify in detail how such assistance systems and similar medical devices perform after post-market training.

This type of system has high levels of autonomy and cannot currently be applied to diagnostic assistance because state-of-the-art computer technology has not yet reached a high enough level to develop these kinds of systems for medical devices. However, future implementations of this type of system would obviate the need for technicians or doctors to create data for post-market training, thereby allowing assistance systems to be more autonomous than those with supervised learning. Moreover, it is also anticipated that innovative assistance systems will be invented that collect and use clinical data for post-market training autonomously, thereby completely obviating the need for intervention by technicians or users for post-market training. Quality control of such autonomous assistance systems, which will have a low degree of or no intervention by humans at all, will be especially important. In addition, the increase in their autonomy will result in much higher requirements for ensuring the performance and risk management of the systems. To manage the performance and risk of the systems, various requirements aforementioned in previous sections should be satisfied at very high levels. Since it may be difficult to specify requirements for such assistance systems in detail, consultations with the Pharmaceuticals and Medical Devices Agency are indispensable prior to applying for their market approval.