Guidance on Clinical Trials Design of Artificial Intelligence(AI)-based Medical Devices

July 2022

MINISTRY OF FOOD AND DRUG SAFETY
National Institute
of Food and Drug Safety Evalution
Medical Device Evaluation Department

Revision History

Guidance on Clinical Trials Design of Artificial Intelligence(AI)-based Medical Devices

No. of Establishment/Amendment	Approval Date	Description
Guidance-0818-01	2017.12.20	Initial establishment
Guidance-0818-02	2019.10.30	Revised to add terminology for retrospective clinical trials
Guidance-0818-03	2022. 7.04	Harmonization with IMDRF/AIMD WG/N67

This guidance explains the scope and describe the position of the Ministry of Food and Drug Safety regarding the review and approval process for the clinical trials design of machine learning-enabled medical devices.

This guidance does not establish legally enforceable responsibilities. Please note that, despite some terms contained herein (such as “shall” or “should”), you are not required to comply with this guidance. In addition, this guidance was prepared based on scientific and technical facts and statutes are valid and effective as of July 2022. The provisions in this guidance is subject to change based on the revision of the relevant statutes or specific factual basis.

※ Guidance for Industry refers to the description of legislation or administrative rules offered to the industry to aid their understanding or the proclamation of the stance of regulatory authority in relation to specific civil affairs (Article 2 of the Regulations on the Management of Guidance, etc. of the Ministry of Food and Drug Safety).

Contact
Digital Health Devices Division, Medical Device Evaluation Department, National Institute of Food and Drug Safety Evaluation, Ministry of Food and Drug Safety
Tel : +82-43-719-3982
Fax: +82-43-719-3940
email: digitalhealth@korea.kr

Contents

I. Overview

1. Background and objective

The "Guidance on the Approval and Review of Artificial Intelligence (AI)-based Medical Devices (for Industry)" introduced the application of retrospective studies in addition to prospective studies for clinical validation of machine learning (ML)-enabled medical devices (MLMD).

Retrospective studies can generate results quickly and efficiently, resulting in shorter clinical trials and lower costs. However, because retrospective clinical trials use existing medical data, trials should be designed considering various issues, such as collecting the clinical trial dataset and establishing the primary endpoints.

Therefore, this guidance is aim to provide information on matters to be considered when designing a retrospective clinical trial for MLMD for public convenience and transparency in the approval and review process.

2. Scope

This guidance applies to the regulatory pathways of clinical trial approval, manufacturing, and import licensing (subject to clinical trial data submission) or technical documentation review (subject to clinical trial data submission) for MLMDs that analyze medical data to diagnose, control, or predict diseases.

3. Definitions of terms

Artificial Intelligence

A computer that uses logic, decision trees, ML, or deep learning to perform tasks that mimic human abilities, such as language comprehension, object and sound recognition, learning, and problem solving.

Blind

Methods to reduce bias caused by investigators, evaluators, and participants learning about the study include single- and double-blind studies.

Crossover Design

Applying the test and control devices to a single subject. The test device is first used in the test-first group, and the control device is used after a specific period.

Equivalence Trials

A study that verifies if the effectiveness of an investigational medical device is equivalent to that of a control medical device.

Human Subject

Individuals from whom the test dataset used in clinical trials is collected.

Institutional Review Board, IRB

An independent body reviewing clinical investigations to protect the rights, safety, and well-being of human participants in a clinical investigation.

Machine Learning-enabled Medical Devices, MLMD

A medical device that uses ML, in part or in whole, to achieve its intended medical purpose.

Non-Inferiority Trials

A study to verify the effectiveness of an investigational medical device is not inferior to that of a control medical device.

Parallel Design

A clinical trial design dividing participants into test and control groups by assigning them to different treatments (two or more groups).

Power

The probability of making the correct decision to reject the null hypothesis if the alternative hypothesis is true.

Primary Endpoint Criteria

In a clinical study protocol, the planned outcome measure is most important for evaluating the effect of an intervention or treatment. Most clinical studies use one primary outcome measure; however, some use more than one outcome measure.

Prospective Study

A method for tracing changes for a specific period after preset factors (risk factors) are studied by observing the changes caused by risk factors.

Randomization

The process of assigning subjects to the investigational medical device or control groups using an established recognized statistical method using an element of chance to determine the unforeseeable assignment in order to reduce bias.

Reference Standard

An objectively determined benchmark that is used as the expected result for comparison, assessment, training, etc.

Retrospective Study

This is a study method without direct contact with the study subjects. This clinical trial is conducted to verify the safety and effectiveness of medical devices using the medical data of subjects obtained through previous medical care or clinical trials rather than by recruiting subjects.

Note: Medical records, medical images, vital signs, pathological examinations, genetic information of subjects, and clinical trial results can be used in retrospective studies.

Superiority Test

A test that verifies if the effectiveness of an investigational medical device is superior to that of a control medical device.

Test Dataset

A set of data that is never shown to the ML training algorithm during training, that is used to estimate the ML model's performance after training.

Training Dataset

A set of data that is used to train the ML model, which is not part of the test dataset.

Ⅱ. Considerations for Clinical Trials Design for MLMD

1. Selection of test dataset

a) Selecting the method and number of test datasets

A test dataset refers to the medical data collected for a clinical trial, such as medical records or patient data from a completed clinical trial. The quality, collection method, and type of the test dataset can affect the assessment of clinical effectiveness in a retrospective trial. Thus, selecting the test dataset is a crucial process. The inclusion and exclusion criteria should be clearly defined during clinical trial design. In addition, the data should be selected based on these established criteria.

The inclusion and exclusion criteria should align with the indications for the medical device and accurately reflect the target population, including disease groups, disease frequency, gender, and other relevant factors.

It is also recommended that the test dataset should be independent of the training dataset used during the development process of the medical device and should be randomized from the collected dataset to avoid bias.

The number of test datasets can be determined using appropriate statistical methods for retrospective clinical trials that consider factors, such as target disease, trial purpose, endpoints, power, and other relevant factors.

The number of test datasets may vary depending on the primary endpoint type, expected clinical outcome, comparison method, and null hypothesis. Based on these factors, an appropriate formula must be developed to calculate the required sample size.

b) Informed consent

Before commencing a clinical trial, informed consent should be obtained and documented from human subjects according to relevant regulations, such as the Standards for Conducting Clinical Trials on Medical Devices (Article 24, Paragraph 1, Item 4 of the Enforcement Rules of the Medical Device Act). The Institutional Review Board (IRB) approves the informed consent form, subject manual, and any other relevant documentation. However, in retrospective trials that use existing medical data, obtaining informed consent may be impractical, or the risk to subjects may be very low. Therefore, waiving for an informed consent in retrospective studies requires careful consideration and is subject to IRB approval.

2. Clinical trials method

a) Establishing the reference standard

A reference standard is the outcome of diagnosing the presence or absence of a disease or a specific condition to be diagnosed or predicted. This standard can be established through single or multiple tests or methods and may include a clinical follow-up of subjects. Table 1 lists examples of reference standards relevant to retrospective clinical trials of MLMD (software). Establishing a reference standard using the data determined by a group of clinical experts is possible when there are no 1) explicit diagnostic criteria for the disease, 2) test results using a reference standard medical device, or 3) confirmed data using the gold standard test method for the disease.

If a group of clinical experts establishes the reference standard, multiple clinicians with appropriate specialties and experience in the indication of the medical device and purpose of the clinical trial should participate.

Clinicians should use the same practice guidelines to identify the disease and resolve disagreements through an appropriate consensus. In addition, the views of clinical societies and organizations related to the disease may be considered if necessary.

To minimize bias in clinical trials, it is recommended that clinicians who participated in establishing the reference standards will not participate as clinical trial evaluators.

Table 1. Examples of reference standards for MLMD.

■ **Explicit criteria for determining disease.**

- **(Example 1)** For medical devices (software) that analyze vital signs to assist in diagnosing the severity of hypertension, the hypertension level or reading used to diagnose hypertension could be set as the reference standard.

- **(Example 2)** For medical devices (software) that analyze various medical information to assist in the diagnosis of systemic lupus erythematosus (SLE), explicitly defined diagnostic criteria could be set as reference standards.

  ※ The diagnostic criteria for systemic lupus erythematosus are the presence of four or more of the 11 items in the lupus classification proposed by the American College of Rheumatology (ACR).

■ **Test results using reference standard medical devices or confirmed data based on gold standard test methods for diseases.**

- **(Example 1)** For medical devices (software) that detect lung cancer using chest X-ray images, medical images of a patient confirmed by computed tomography (CT) or magnetic resonance (MR) imaging can be used as a reference standard.

- **(Example 2)** For medical devices (software) that aid in the diagnosis of breast cancer using X-ray imaging, medical images of a patient confirmed by biopsy can be used as a reference standard.

■ **Data determined by a group of clinician experts.**

- **(Example 1)** In the absence of a clear diagnostic criteria or other testing methods for a disease, data determined by a group of clinical experts can be set as a reference standard.

- ***What should be included in the reference standards for clinical trial protocols for medical devices?***

- Types of reference standards.
- Requirements for establishing a reference standard and the process of establishing a reference standard.
- Specialization and experience of clinical experts involved in establishing the reference standard.
- If necessary, the views of relevant clinical societies or organizations when establishing the reference standard.

b) Clinical trials design

Clinical trials should be designed according to the purpose of the trial, characteristics of the product, method of proving effectiveness (superiority, equivalence, non-inferiority, etc.), and criteria for evaluating effectiveness. In addition, retrospective clinical trials can apply parallel or crossover designs depending on the purpose of the trial.

Table 2 lists examples of clinical trial designs applicable to retrospective clinical trials of AI-based medical devices (software). However, this is not necessarily limited to the following examples:

c) Multicenter trials

Multicenter trials can provide a good basis for obtaining sufficient trial datasets or generalizing trial outcomes. Multicenter trials are not required for retrospective trials. The decision to perform multicenter trials can be made in consultation with the sponsor and participating sites.

Table 2. Examples of clinical trial design for MLMD

- ■ **Designing a clinical trial to evaluate the diagnostic effectiveness of a medical device.**
  - **(Example 1)** A clinical trial comparing the readings of a medical device with a reference standard in a single arm of a test dataset.
  - **(Example 2)** A clinical trial in which the same test dataset is assigned to the test and control groups. The readings of the medical device in the test group are compared to those of a clinician in the control group.
- ■ **Design of clinical trials to evaluate the improvement of clinicians' diagnosis rate using a medical device.**
  - **(Example 1)** Comparing the accuracy of clinicians' readings before and after using the product with sensitivity, specificity, AUC, etc., by organizing the test dataset into a single arm.

d) Randomization and blinding

Randomization and blinding are significant design methods for minimizing bias in clinical trials. In retrospective trials, the test dataset is selected by randomly drawing from a known population of human subject data. In multicenter trials, when the subject data are selected randomly, the randomization process can be centrally managed, and randomization can be performed separately at each site. In addition, if the trial design involves clinicians in evaluating the test or control groups, they should be blinded to results of the human subject data.

■ ***What should be included in a clinical trials protocol for medical devices regarding randomization and blinding?***

- ■ Randomization and blinding methods used to minimize bias.
- ■ The method and rationale. (If a method other than randomization and blinding is selected to minimize bias)

e) Control group selection and control devices

If clinical effectiveness is to be assessed by comparing the reading accuracy of a medical device with that of a clinician, the clinician may participate as a control group evaluator. It is recommended that clinicians participating in the control group assessment be independent of the expert group of clinicians involved in developing the reference standards.

A control group is not required for retrospective studies; however, if the objectives of the study require selecting a control group, the study can be designed accordingly.

■ ***What should be included in the clinical trial protocol for medical devices considering a control group and device?***

- ■ The selection criteria for clinicians, such as specialty, experience, and number of clinicians. (If clinicians are involved in the evaluation of the control group)
- ■ Regulatory description of the control device being evaluated, including its MFDS approval/certification/listing number, nomenclature, model number, and manufacturer name. (If a control device is set up)

f) Clinician consensus decision

A final disease diagnosis can be made when multiple clinicians establish a reference standard or participate in a clinical trial.

The clinicians' consensus methodology includes procedures for obtaining objective readings and clarifying the basis for each reader's judgment, reasons for any disagreement, and rationale for the consensus. This consensus method is voluntary and can be set by the sponsor, considering the number of readers involved in the trial.

For example, a three-reader study may adopt a two-reader consensus. By contrast, a two-reader study may have a consensus led by a clinician with significant experience in the field.

- ***What should a medical device protocol include about clinician consensus decision?***

> - The medical device protocol should detail the number of readers, method, and procedure for consensus among the studies.

3. Primary endpoint and statistical analysis

a) Primary endpoint

In retrospective clinical trials of MLMD, the clinical primary endpoint could be specified as sensitivity, specificity, ROC curve, AUC, etc., of diagnosis using a reference standard or concordance rate if there is no reference standard. If necessary (e.g., in rare disease trials), positive and negative predictive values based on prevalence can be calculated using the reference standard (Table 3).

Table 4 lists examples of primary effectiveness endpoints in retrospective clinical trials.

Table 3. Description of the major primary endpoint for MLMD.

- **Sensitivity**
  Probability to identify those with the disease among people with the disease.

- **Specificity**
  Probability to identify those without the disease among people without the disease.

- **Positive Predictive Value**
  Fraction of those with a specific disease characteristic among those classified with that specific disease characteristic.

- **Negative Predictive Value**
  Fraction of those without a specific disease characteristic among those classified without that specific disease characteristic.

- **ROC (Receiver Operating Characteristic) Curve**
  The ROC is a graph drawn using sensitivity and false-positive rate (1- specificity) based on diagnostic test results. This curve can be used to assess the diagnostic performance distinguishing positive or negative.

- **AUC (Area Under the Curve)**
  Refers to an area under the ROC curve indicating diagnostic accuracy. In a range of 0.5 to 1.0, the closer the value to 1, the better the performance.

- **Positive Percent Agreement**
  Proportion of diagnostics with positive results that match the control diagnostic with a positive result.

- **Negative Percent Agreement**
  Proportion of diagnostics with negative results that match the control diagnostic with a negative result.

Table 4. Examples of the primary endpoint for MLMD

- **Calculate the sensitivity, specificity, and AUC of a single software arm.**
- **Calculate and compare the sensitivity, specificity, and AUC of software readings with the sensitivity, specificity, and AUC of clinician readings.**
- **Compare the change in sensitivity, specificity, and AUC before and after clinicians use the software.**
- **Calculate the overall agreement, positive percent agreement, negative percent agreement, and Kappa statistics between software and clinician readings in the absence of a reference standard.**

The sensitivity, specificity, positive predictive value, and negative predictive value are compared to the results of the reference standard and calculated in a 2×2 table format (Table 5).

Table 5. Methods for calculating sensitivity, specificity, positive predictive value, and negative predictive value.

Result		**Reference Standard**		
		True	False	Overall
Investigational MLMD	Positive	True Positive(TP)	False Positive(FP)	TP+FP
	Negative	False Negative(FN)	True Negative(TN)	FN+TN
	Overall	TP+FN	FP+TN	TP+TN+FP+FN

- **Sensitivity = 100 × TP/(TP+FN)**
- **Specificity = 100 × TN/(FP+TN)**
- **Positive Predictive Value(PPV) = 100 × TP/(TP+FP)**
- **Negative Predictive Value(PPV) = 100 × TN/(TN+FN)**

Clinical validity can be confirmed by calculating the agreement rate through a comparative evaluation with the clinician's reading results when data determined by a group of clinician experts are used as a reference standard or when sensitivity and specificity are difficult to apply.

The agreement rate does not indicate diagnostic accuracy; thus, it is a lower-level evaluation method compared with sensitivity and specificity analysis. There are no appropriate criteria for the agreement rate; thus, it should be interpreted carefully.

If sensitivity and specificity can be calculated, selecting the agreement rate as an evaluation variable should be avoided as much as possible. Only when sensitivity and specificity are difficult to calculate should the overall agreement, positive percent agreement, negative percent agreement, and kappa statistic be selected as evaluation variables. Table 6 lists the methods used for calculating the agreement rate.

Table 6. Methods for calculating agreement rate.

Result		**Control method**		
		Positive	Negative	Overall
Investigational MLMD	Positive	A	B	A+B
	Negative	C	D	C+D
	Overall	A+C	B+D	A+B+C+D

- ■ **Sensitivity = 100 × TP/(TP+FN)**
- ■ **Specificity = 100 × TN/(FP+TN)**
- ■ **Positive Predictive Value(PPV) = 100 × TP/(TP+FP)**
- ■ **Negative Predictive Value(PPV) = 100 × TN/(TN+FN)**
- ■ **Kappa statistic = 2(A×D-B×C)/[(A+B)×(B+D) ×(A+C) ×(C+D)]**

  Kappa statistic: statistic measure of inter-evaluator agreement or reliability for categorical data.

b) Success criteria

The success criteria for clinical trial evaluation results can be set autonomously by the sponsor, who must provide the rationale and evidence for establishing the success criteria. The success criteria can be set by selecting one or more of the following three types of control groups. Table 7 lists the methods.

Table 7. Methods for calculating the agreement rate.

- ■ **Application device & control device**: establish target performance based on the performance of licensed medical devices.
- ■ **Application device & historical control group (scientific/clinical paper):** established by citing the clinical success criteria of the paper.
- ■ **Application device & standard diagnostic method (doctor's practice):** establish standard diagnostic methods based on papers, etc.

The established success criteria should be described in the performance and method of use of the technical documentation submitted for approval by the Ministry of Food and Drug Safety and should be specified in the user manual such that users can check it.

III. References

1. Guidance on the Approval and Review of Artificial Intelligence-Based Medical Devices (for Industry), Ministry of Food and Drug Safety, 2022
2. Guidance on Pharmaceutical Clinical Trial Statistics, Ministry of Food and Drug Safety, 2016
3. Guidance on Clinical Performance Review of In Vitro Diagnostic Medical Devices, 2015
4. Clinical Research in a Nutshell: Everything You Need to Know About Designing and Executing Clinical Trials, Clinical Trials Review and Analysis Society, 2015
5. Statistical Guidance on Reporting Results from Studies Evaluating Diagnostic Tests, FDA, 2007
6. Use of Real-World Evidence to Support Regulatory Decision-Making for Medical Devices, FDA, 2017
7. Machine Learning-enabled Medical Devices: Key Terms and Definitions, IMDRF, 2022
8. ISO 14155 : Clinical investigation of medical devices for human subjects — Good clinical practice, ISO, 2020
9. Clinicaltrials.gov

Digital Health Devices Division
Medical Device Evaluation Department
National Institute of Food and Drug Safety Evaluation
Ministry of Food and Drug Safety

Email: digitalhealth@korea.kr

https://www.mfds.go.kr/eng/medicaldevice.

MINISTRY OF FOOD AND DRUG SAFETY
National Institute
of Food and Drug Safety Evalution
Medical Device Evaluation Department