

Provisional Translation (as of October 2026)*

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

Review Points for Medical Devices Utilizing Generative AI

(Draft)

Pharmaceuticals and Medical Devices Agency

Center for Product Evaluation

September 14, 2026

Table of Contents

1. Introduction	3
2. Scope of These Review Points.....	4
3. Definitions	4
4. Characteristics of GenAI	5
4.1. Distinctive Technological Characteristics of GenAI	5
4.2. Possible Features of GenAI-Enabled Medical Devices.....	8
4.3. Development and Operation Models Assumed for GenAI-Enabled Medical Devices.....	8
5. Challenges in Regulatory Evaluation and Operation of GenAI.....	9
5.1. Challenges in Evaluation and Operation	9
5.1.1. Uncertainty and Variability of Outputs	9
5.1.2. Evaluation of the Appropriateness of Outputs	10
5.2. Challenges in Use	10
5.3. Other Matters	12
6. Principles on Evaluation.....	12
6.1. Principles on Overall Evaluation	12
6.2. Premarket Review.....	13
6.2.1. Usefulness.....	14
6.2.2. Product Performance.....	14
6.2.3. Other Evaluations	19
6.3. Post-Marketing Measures	19
6.3.1. Post-Marketing Reporting.....	19
6.3.2. Design and Operation based on Human Factors.....	19
7. Conclusion	21
Appendix 1: Examples of Development Models	22
Appendix 2: Technological Characteristics that may Raise Issues and Suggestions for Development.....	23
Appendix 3: Positioning and Limitations of Benchmark Testing	27

1. Introduction

In recent years, there has been a rapid increase in public interest on Generative Artificial Intelligence (GenAI). GenAI is "a generic term for AI that can generate texts, images, software, or other formats." In particular, GenAI in recent years has been trained on extremely large datasets, enabling it to generate diverse answers and creative content not dependent on a specific task in response to a wide range of user inquiries through natural language texts or speech inputs. GenAI is built as a Large Language Model (LLM) or a Vision Language Model (VLM) and is expected to be used widely in the medical sector, including clinical support, documentation, and image analysis. However, from a regulatory perspective, no sufficient framework has yet been established for the evaluation and risk management of GenAI technologies. Although AI has been discussed extensively^{i,ii,iii,iv}, most of these discussions have addressed Discriminative AI, which is specific to a particular task, rather than GenAI. Unlike conventional Discriminative AI, for which the target function (task) can be clearly defined, and output errors can be evaluated quantitatively and objectively using rule-based criteria through inference outputs such as labels, GenAI is highly generalized and versatile, making it difficult to limit the scope of use. Judging output errors in natural language text or in another format leaves room for interpretation. As a result, limited information is available to clarify, in a systematic and practical manner, questions such as which aspects of GenAI are regulatory challenges and which characteristics of GenAI should be recognized as specific risks in regulatory development.

The objective of these review points is to comprehensively summarize the above, based on current knowledge, and to discuss regulatory principles, including evaluation methodologies and post-marketing measures. However, it should be noted that at the time of publication of this document, no GenAI-enabled medical device has yet been reviewed for approval by PMDA. Therefore, this document summarizes a draft of review points considering the current principles of approval review. The aim is to gather a wide range of opinions based on this document and refine it to a more appropriate set of review points.

Careful discussions are being held on the handling of GenAI-enabled medical devices, including the extent of items to be handled as medical devices (i.e., qualification as a medical device). These review points target products that qualify as medical devices, including their claims, specified in the Act on Securing Quality, Efficacy and Safety of Products Including Pharmaceuticals and Medical Devices (hereinafter referred to as "Pharmaceuticals and Medical Devices Act"). These review points may need to be adjusted if determination on whether a product qualifies as a medical device is updated in the future. Since the development of AI technology including LLMs is remarkable, this document should be revised as appropriate in a more rational manner based on future experience and knowledge.

Generally, approval review of medical devices is based on the diversity of each product. Even for products that are identical, including in terms of clinical positioning, different evaluation methodologies may be acceptable if they are reasonably justified. Note that the descriptions in these review points are provided only as examples, and flexible decisions should be made for each individual case.

2. Scope of These Review Points

The concept of "GenAI" is broad, and it is generally understood as "a generic term for AI that can generate texts, images, software or other formats." These review points apply to GenAI that is implemented to generate diverse answers and creative content, that is not dependent on specific tasks, in response to a wide range of user inquiries through natural language text or speech inputs (hereinafter referred to as simply "GenAI", unless otherwise explained). Specifically, such GenAI are assumed to be implemented as an LLM or a VLM.

In addition, medical devices with functions enabled by the GenAI described above (hereinafter referred to as "GenAI-enabled medical device") are covered by these review points.

3. Definitions

- Development concept: Background or history that led to development of the product under application, natural history (presence of existing therapies and treatment process with existing therapies) or actual medical conditions (clinical performance) of the target disease, severity of the disease, current development in Japan and overseas and other elements.
- Product under application: A product seeking marketing approval. The term "product under application" is intended for marketing approval applications, and "product under consultation" is intended for PMDA Consultation.
- Product performance: Processing performance and capacity to the input data in the product under application.
- Design concept: Functions and level of performance of the product intended to be developed based on the development concept.
- Authority: Regulatory authority. Pharmaceuticals and Medical Devices Agency (hereinafter referred to as "PMDA") for review and consultation.
- Clinical positioning: By whom, and for what purpose the product under application will be used in clinical practice.
- Clinical utility: The value of using the product under application in clinical practice and patient outcomes.
- Broadly defined GenAI: A generic term for AI that can generate texts, images, software or other formats.
- GenAI (subject to these review points): Within the broadly defined GenAI, those capable of generating diverse answers and creative content not dependent on a specific task in response to a wide range of user inquiries through natural language texts or speech inputs.
- GenAI models: Deep neural networks trained on data that output text, images, software, and other content in response to natural language texts and speech inputs.
- GenAI systems: Systems built around a "GenAI model" that generates responses and in which multiple components operate in a coordinated manner, such as an "interface" through which the user interacts with the system, "guardrails" that process inputs and outputs safely, and a "toolset" that retrieves information and performs other processing.
- GenAI-enabled medical devices: GenAI systems that qualify as a medical device under the Pharmaceuticals and Medical Devices Act.

- GenAI functions: Functions of GenAI-enabled medical devices that are achieved using GenAI models at their core.
- Prompt: A statement instructing a GenAI system what to do.
- GenAI content: Texts, images, software or other formats generated by GenAI models or systems, which have meaning that is valuable as information when interpreted by the user.
- False information: Fabricated information that is not based on facts.
- Erroneous information: Information contrary to the facts.
- Harmful information: Information that directly harms the life or safety of patients, such as recommending inappropriate self-treatment or instigating dangerous behavior.
- Inappropriate guidance: Output that discourages patients from receiving necessary medical care, such as output that deters them from seeking medical consultation or recommends alternative treatment without scientific evidence.

4. Characteristics of GenAI

4.1. Distinctive Technological Characteristics of GenAI

The distinctive technological characteristics of GenAI and the related points to be noted are summarized in terms of five aspects^v.

(1) Large-Scale Training Processes and Model Architecture

GenAI models are known to have scaling laws in which their performance improves as the size of the model and the training dataset increases. Therefore, for example, large neural networks with hundreds of billions of parameters are built and pre-trained using the vast and diverse data that exist on the Internet and elsewhere. The model then undergoes post-training, using methods such as supervised fine tuning and reinforcement learning, to enable appropriate and desirable responses to user inquiries.

The development and operation of such large-scale models typically needs enormous computational resources and large investments. For this reason, many medical device vendors, as developers, are expected to use external foundation models and to be actively involved in downstream processes, such as system integration and interface optimization, rather than developing and operating in-house GenAI models from scratch.

[Points to be Considered]

Large-scale training processes and model architecture may make reliance on third parties practically unavoidable in the development, and even in the operation of the GenAI model, which is the core component of the medical device. Details of the dataset used in the development process may not be disclosed to the developers, causing challenges such as difficulty in identifying the root causes of problems, obscuring of potential dataset biases, and difficulty in justifying appropriateness of test data in the evaluation process considering data leakage (a problem in which the test data is included in the training dataset).

(2) Content Generation Capability According to Context

While conventional Discriminative AI presents fixed outputs, such as labels, GenAI models generate content that is subject to human interpretation, such as texts, images, and software, depending on prompts entered by the user. Such content may be generated by adding randomness to produce fluctuating output for the identical prompt.

In addition, such content varies more adaptively depending on the context, such as interaction logs, as well as the prompts that the user enters. This allows for two-way and continuous interaction with users, and outputs from GenAI models also reflect more user-specific contexts.

[Points to be Considered]

There are risks of hallucination including generating content with false information, that seems true but is not based on facts, and erroneous information. This raises the challenge of how to ensure the accuracy of the output as a medical device, and whether the authenticity of the generated output can be properly interpreted by typical users. In addition, small differences in the expression of inputs can alter the generated output, which raises concerns about dependence on user's input skills and reduced reproducibility. In addition, accumulation of conversation history is associated with risks of gradual changes and bias in model behavior.

(3) Generalization and Versatility

While conventional Discriminative AI specializes in single tasks, such as presenting output to a given operation to accomplish a particular objective (e.g., detecting defined abnormal findings from medical image inputs), GenAI provides generalization and versatility. GenAI can seamlessly execute basic tasks, such as question answering, summarization and translation, to advanced tasks, such as diagnostic support for multiple diseases, through interactive instructions to a single model.

These functions are activated dynamically in response to user prompts. In addition, even if a function is difficult to perform with high accuracy under standard conditions, it is possible to adapt to new tasks without parameter updates by including examples of the relevant task in the prompt (in-context learning). Alternatively, parameters may be adjusted by fine tuning or the like to obtain specific functions after deployment.

[Points to be Considered]

The generalization and versatility of model functions enable multifunctionality in the medical device, such as diagnostic support for multiple diseases. However, it complicates product performance evaluations. In addition, the functional boundaries need to be clarified since tasks other than the claimed function approved as a medical device may be executed. Therefore, to ensure that tasks unintended by the applicant will not be executed, it is necessary to implement a guardrail in the model to prevent off-label use (use outside of the approved intended use or population), inappropriate use (use against precautions or recommended procedures), or misuse (unintended use such as operational errors or misinterpretation of outputs).

(4) Interoperability with External Systems

GenAI models, such as LLMs, not only operate independently but are also advancing towards enhanced interoperability by linking with external systems and data sources. Examples include retrieval-augmented generation

(RAG), that retrieves recent medical papers from external databases and incorporates them into the response, and functions that call external tools, such as calculation functions, databases, and other systems, and incorporate the resulting outputs into the model's own generation process. This makes it possible to dynamically refer to the latest and specialized external information resources rather than being closed to the internal knowledge of the model itself.

[Points to be Considered]

Interoperability with external systems increases the convenience and accuracy of GenAI models, while raising the challenge of strong environmental dependencies. There is a risk that the performance as a medical device may be affected if the referenced external data is changed or deleted, or if the connected system fails. Accordingly, it is necessary to clearly control the scope, version, and availability of the connected system and to establish a framework to monitor and detect changes in performance caused by external dependence.

(5) Autonomy of Actions

Based on connection capabilities with external systems described in (4), GenAI models may be capable of acting autonomously, in which they develop their own procedures to achieve a given goal and execute across a sequence of steps. One example is agentic AI functions that can repeat the loop of identifying a task from the input context, breaking it down into subtasks, designing a plan, calling appropriate external systems as tools, observing and evaluating the result, and determining the next action without human sequential instructions.

Such autonomy is not binary but can be seen as continuous phases that align with the degree of human involvement, like the levels of automation (levels 0-5) in autonomous vehicles. In other words, it can be assumed that there is a range from a low level at which humans execute and approve each action (humans as operators), to an intermediate level at which humans review and approve model proposals, to a high level at which humans only monitor the results where models continuously execute a series of actions for a specified goal (humans as observers). Even for models with the same performance, the level of autonomy at the time of operation is a design choice independent of the level of model capability. High-performance models can be operated at a limited low level, or vice versa.

[Points to be Considered]

While the autonomy of actions makes it convenient to perform a series of complex clinical tasks, it also raises challenges in controllability of the entire system. While it is necessary to clarify the boundary between the scope of functions that can be executed autonomously and functions that require human supervision and intervention, it is also important to note that as autonomy increases, GenAI models can find loopholes in constraints and guardrails implemented by humans and adopt unintended workarounds or deviations to optimize the achievement of a given objective (reward hacking). In addition, there are risks of amplifying errors from hallucinations described in (2) and off-label tasks described in (3) linked and propagated to subsequent actions in these autonomous loops without human confirmation. Furthermore, since action sequences are determined dynamically in response to prompts and external environments, their behavior is difficult to predict comprehensively in advance, complicating performance evaluations and assurance of reproducibility. Therefore, implementation of human confirmation, stoppage, and intervention into the design (e.g., human-in-the-loop), establishment of a framework to follow-up and audit executed

action sequences, and clarification in advance of responsibility for events that may occur according to the expected level of autonomy would be needed in medical device risk management.

4.2. Possible Features of GenAI-Enabled Medical Devices

The combination of these properties allows medical devices to be realized with the following features:

- Able to answer user's questions and provide necessary analysis results by generating flexible responses to a variety of questions from users. In addition, the user's conversation history is accumulated to provide optimized output according to the user's skill and level of knowledge. Such automatic adjustments allow users, who do not have specialized medical knowledge, to achieve advanced functions, such as treatment support.
- Have functions that analyze, in an integrated manner across modalities, not only texts but also information contained in various types of modality data, such as images and tabular data, in response to inquiries from users using natural language texts. This enables functions and performance that could not be achieved in the past, such as providing diagnostic performance superior to that of conventional medical devices, which can handle only certain modalities, or interactively supporting reader interpretation of radiographic images in natural language texts.
- Have functions that respond to inquiries from users based on the latest findings and analyze them through real-time linkage with external information processing systems, such as medical literature databases. In addition, when agentic AI functions are implemented, the device may autonomously select the most appropriate source from among various external information systems, observe and evaluate the results obtained, and then present the results of the final integrated analysis to the user.

4.3. Development and Operation Models Assumed for GenAI-Enabled Medical Devices

Since significant investment is necessary to research and develop GenAI models from scratch, many foundation model developers provide highly generalized models as cloud APIs or as open-weight models. Accordingly, development is expected to be divided, with GenAI models provided by external developers and the applicant optimizing the entire GenAI system, including the model concerned, to design and develop elements (e.g., input/output design for clinical use, guardrails, test datasets and procedures, operation and monitoring systems) that meet medical device requirements (e.g., intended use, clinical positioning, quality management, risk management, verification and validation, change controls).

For example, the areas directly controlled by applicants and those dependent on external developers vary according to the development model (e.g., whether the model can be updated or retrained, possibility of obtaining training data, locking inference environments, security requirements, log availability). Examples include use of proprietary models via cloud API and other methods, acquisition and operation of open-weight models in the applicant's own environment, or conduct of additional training, such as fine-tuning, for medical use (see [Appendix 1](#)). Challenges in evaluation or use that need to be addressed, as well as accountability and management policies required of applicants (including contracts and Service Level Agreements) should also be specified according to the development model. Even for identical products, appropriate risk management and software life cycle activities should be implemented

depending on the development model.

5. Challenges in Regulatory Evaluation and Operation of GenAI

Approval review of medical devices under the Pharmaceuticals and Medical Devices Act requires evaluation of the efficacy and safety of the product under application to be marketed (a guidance has been published for the general development concept of Software as a Medical Device). "Effectiveness" can be categorized into "usefulness" and "product performance." Usefulness refers to the contribution to clinical outcomes by the product under application, such as improved diagnostic accuracy and treatment performance. Product performance refers to the performance of the product under application under specific conditions.

For example, even if the medical device is realized using a GenAI model, it is likely that it will follow conventional medical device evaluation policies if the use of the device is limited, as with conventional medical devices (e.g., detecting XX findings on an image, presenting risk scores for disease YY), and the procedure for generating outputs is fixed.

This section broadly categorizes and organizes the challenges related to the evaluation and use of medical devices from technological aspects, reflecting the characteristics of GenAI described in Section 4, as follows. While ethical challenges are also significant in relation to GenAI, they are not addressed in these review points, which focus on considerations for approval review. Challenges on other applicable legislation are recognized but also excluded, such as the Act on the Protection of Personal Information with respect to the handling of medical information, and the Copyright Act with respect to copyrighted works included in training data.

- Challenges in evaluation and operation: Challenges arising from technological characteristics
- Challenges in use: Challenges relating to proper use by users
- Ethical challenges: Challenges relating to matters such as responsibilities (not addressed in these review points)
- Legal challenges: Challenges relating to other applicable legislation, such as the Act on the Protection of Personal Information and the Copyright Act (not addressed in these review points)

5.1. Challenges in Evaluation and Operation

Technological characteristics that may raise issues in GenAI-enabled medical device evaluations are summarized in [Appendix 2](#). The challenges in evaluation and operation arising from the above are summarized to two points. They are described in this section, respectively.

- Uncertainty and variability of outputs
- Evaluation of the appropriateness of outputs

5.1.1. Uncertainty and Variability of Outputs

The outputs of GenAI models may vary with slight differences in user inputs and in context, such as interaction logs. In addition, some GenAI models, such as LLMs, are designed to generate expressive outputs by sampling with certain randomness rather than constantly selecting the most probable token. When they are connected to external systems or are using proprietary models provided by external developers via APIs, they are also affected by the external

elements, such as updates in external environments and foundation models. In addition to fluctuations in the level of output, excessive application of individual optimizations, such as adjustments per installation site and interaction logs, may lead to the use as a medical device that was not intended by the applicant. This makes it a challenge to verify the reproducibility of the outputs of GenAI models.

In addition, in conventional reviews, medical devices are generally required to evaluate the efficacy and safety within a specific range based on its predefined range of variability (e.g., range of input conditions). However, for GenAI-enabled medical devices, it is extremely difficult to define its "specific range", even when a certain level of control will be provided, due to diverse input ranges and strong effect of uncertainty and variability in outputs as described above.

Improvement Design within Approval for Timely Evaluation and Notice (commonly known as "IDATEN") is a system under the Pharmaceuticals and Medical Devices Act corresponding to the plasticity of machine learning. IDATEN is useful in providing timely updates to specific changes, but there are limitations to its ability to accommodate products with characteristics that change in real time. There have been cases of agentic AI functions attempting to work around human-designed safety mechanisms, and the scope that can be controlled and managed in advance may be limited. Therefore, the conventional framework to "ensure sufficient efficacy and safety solely with premarket evaluations" is difficult to apply. To ensure efficacy and safety of products that change in time, it is essential to actively implement the rebalancing framework, which covers efficacy and safety evaluation before and after marketing.

5.1.2. Evaluation of the Appropriateness of Outputs

The challenge in evaluating GenAI-enabled medical devices is to determine the appropriateness of its outputs with reproducibility and objectivity. Appropriateness includes the "accuracy of the information", whether the generated information is consistent with medical facts and evidence, as well as the "suitability to context", whether the generated response is in line with the clinical background of the patient, the user's intent, and use conditions. As for the former, since false or erroneous information by hallucination can be included, the challenge is to develop a methodology to measure its tendency objectively. For the latter, since there are many tasks for which no single correct answer exists, the challenge is to establish criteria to determine acceptability.

Moreover, because GenAI outputs have clinical value only when the user interprets its content, appropriateness is not determined by the output alone but depends on the user's interpretation. Therefore, it is necessary not only to evaluate the content of the output, but also to consider how the intended user understands the output and how it leads to their judgment and behavior. In addition, these decisions on appropriateness are not uniform across patient populations and use conditions, and performance may vary in specific populations and circumstances, depending on the inherent biases in the training data. Therefore, identifying the assured scope and limitations of appropriateness is an essential challenge in premarket evaluation, not only by confirming the mean performance but also by identification of clinically significant populations and conditions.

5.2. Challenges in Use

This section summarizes the challenges relating to the appropriate use of GenAI-enabled medical devices by their

users, based on the features mentioned above.

(1) Explainability

GenAI and other deep neural networks produce outputs through multilevel parameter weighting and patterning rather than through explicit logic accumulation. Therefore, in principle, it is difficult to logically explain to users the reasoning process by which the model led to its conclusion. As a result, for example, if decisions regarding patient care were made based on GenAI outputs without clear supporting evidence, physicians or other users may not be able to demonstrate sufficient accountability to patients and consequently may be unable to obtain appropriate informed consent. Therefore, to improve the explainability of the system, GenAI-enabled medical devices should preferably be designed to enable users to verify and understand the output as far as possible. One example is to present the output together with the source on which it is based, such as medical literature.

(2) Off-label Use, Inappropriate Use, or Misuse

When GenAI systems are designed and operated to answer any questions from the user, there are risks of deviations regardless of the user's intentions. These include "off-label use" that request tasks outside the scope of approval, "inappropriate use" that violates precautions and recommended procedures, and "misuse" due to unintended operational errors.

In addition, GenAI models may generate outputs beyond the scope of their intended functions, such as assumed tasks and clinical positioning, without the user being aware of it, owing to contextual changes such as the accumulation of conversation history during operation. Such deviations in output are also associated with the risk of inducing significant "misuse" which may lead users to misinterpret it as valid medical information.

(3) User Literacy

GenAI content leaves the judgment of authenticity and usefulness to the user's own interpretation (interpretation dependency). However, GenAI models carry the risks of "hallucination", generating false or erroneous information that is not based on facts as if they were true. It is extremely difficult to fully detect and eliminate such false and erroneous information in advance only with technological measures.

If the system generates plausible false or erroneous information, where the judgment of its output depends on the user's interpretation, the risk of the user ignoring the error without noticing increases. Therefore, users will need to have a certain level of literacy that enables them to critically verify and judge AI outputs with their own expertise, rather than accepting AI outputs uncritically.

(4) Cognitive Bias

While the "data bias", explained previously, is a system bias inherent in AI models due to training data, the "cognitive bias" refers to a user's psychological and cognitive characteristics of receiving AI outputs, i.e., a bias in perception due to "human factors".

In particular, there is a risk that cognitive bias rooted in this human factor may manifest itself in chatbot-type treatment support products, where the patient themselves is the user. For example, users may overestimate advice

of GenAI owing to “algorithmic authority (tendency to prioritize the output of products over the output of human judgment)”, thereby damaging relationships with physicians. In addition, “automation bias” should be considered, where users overestimate and rely on the outputs of algorithms such as GenAI as these appear more objective and accurate than human judgment.

5.3. Other Matters

In addition to assessing the quality, efficacy, and safety of the product under application, its acceptability to clinical practice should be considered. In the following cases, development should proceed with a view to consulting relevant academic societies and creating necessary usage guidelines.

- Where GenAI is not merely assistive but can also complete or automatically execute sequential clinical workflows as an autonomous agentic AI, without the intervention of a physician or other healthcare professionals.
- Where GenAI makes comprehensive determinations based on a vast amount of multimodal information that humans cannot process at the same time, such as texts, images, and genetic information, making it extremely difficult for users to intuitively understand and verify the basis of their outputs.
- Where AI chatbots convey a high degree of empathy and confidence toward patients, which may lead to patients prioritizing AI responses over physician consultations, thereby damaging traditional physician-patient relationships.

6. Principles on Evaluation

6.1. Principles on Overall Evaluation

Conventional reviews have focused on premarket evaluation of quality, efficacy, and safety, while post-marketing data collection and measures were implemented in a supplementary manner. However, as described above, the following characteristics of GenAI-enabled medical devices make it difficult to implement sufficient premarket evaluation.

- Uncertainty of outputs
- Changes in state
- Evaluation of the appropriateness of outputs in context
- Impact of user acceptance
- Challenges in use

Based on the above, the concept of the "Total Life Cycle Approach," which consistently manages quality and risk throughout the entire product life cycle, becomes more important. In other words, it is necessary to identify the properties that should be subject to premarket evaluation, and to determine the extent to which those properties should be maintained through continuous post-marketing management.

Therefore, it may be acceptable to conduct premarket evaluations of clinical efficacy and performance under specific conditions representing intended use (including the extent to which efficacy and performance are maintained, as well as confirmation of conditions in which they fail), provided that, in post-marketing, the premarket evaluation status is

maintained and a system is established to actively identify incidents and respond appropriately when they are discovered.

[Reference] Usefulness of Sharing Risk Management Activities between Developers and Reviewers

Medical devices are designed and evaluated according to risk management principles (e.g., JIS T 14971/ISO 14971). This is equally important for GenAI-enabled medical devices. Whether the characteristics of GenAI described in Section 4 become hazards to the product under application as a medical device; how the risks associated with the hazards are controlled and in acceptable levels; and, if the risks are not controlled to an acceptable level, how to implement risk assessment of the product under application, including its benefits and post-marketing management - sharing the set of risk management activities promotes the reviewer's understanding of the design of the GenAI system, including its guardrails and other components, and of the functions and performance to be evaluated.

Sharing information between reviewers and developers on how risk management was implemented in the development of the product under application is useful in conducting medical device reviews. As mentioned earlier, GenAI-enabled medical devices vary in their outputs due to various factors, which may not only affect accuracy and suitability of the output but also cause functional changes. Some of these variation factors may be difficult for the applicant to control, such as log data documenting the operation of the product under application, and external reference databases that are updated in real time, such as the internet. Dynamic variation factors outside of the applicant's control may cause changes over time in the performance and function of products after marketing. Through risk management processes and decisions based on the process, appropriate mutual understanding of the characteristics of GenAI to be considered and its risk levels will be obtained, enabling sufficient discussion in the review.

6.2. Premarket Review

This section describes the assessment policy for premarket reviews. The following items presuppose comprehensive consideration of the hazards arising from the characteristics of GenAI described in Section 4. However, it is appropriate to consider and select the necessary items based on the understanding of the product under application. Assuming that post-marketing management will be implemented, premarket evaluations should focus on the following points:

- Usefulness of outputs at a given point in time and under specified conditions (see 6.2.1)
- Product performance of outputs at a given point in time and under specified conditions (see 6.2.2)
- Stability and reproducibility of product performance of outputs across different times and under different conditions (see 6.2.2)
- Ensuring the absence of harmful outputs (see 6.2.2)

The above list is an example of premarket evaluation concept as of today. It does not preclude more reasonable evaluation methodology based on the characteristics of the product under application.

6.2.1. Usefulness

Considering the range of GenAI-enabled medical device developments covered by these review points, a product under application intended to contribute to both treatment and diagnosis is expected to be developed. In such cases, usefulness should be evaluated for both treatment and diagnosis.

Two approaches to evaluate usefulness are expected: evaluating whether clinical outcomes are improved using the product under application (e.g., improvement of treatment outcomes and diagnostic performance), and examining whether the product performance results justify clinical significance. The approach to be adopted should be determined on a product-by-product basis based on the sufficiency of the data on the product under application and its clinical positioning.

If the usefulness of the product under application needs prospective evaluation, a clinical trial is expected. Especially in GenAI-enabled medical devices for diagnostic support, a product under application that is not limited to a specific disease may be developed (e.g., products intended to support medical consultation in internal medicine). Ideally, the usefulness of the product under application should be evaluated across the situations in which it can be used, but careful consideration should be given to whether feasible test protocols can be established for the product under application without limiting the target disease or symptoms. A policy to gradually expand claims from what can be implemented should also be considered (e.g., gradually expanding the intended use after obtaining initial approval in narrowly defined intended use).

When preparing test datasets, the type of data and the definition of correct answers need to be justified. However, it may be technically difficult to fully verify the clinical utility of a product under application, such as coverage of all prompts that can be entered by an individual user in the actual use, the suitability of the responses to the individual contexts, and the independence of the test data due to the confidentiality of training data, such as when using third-party foundation models. Therefore, to verify usefulness, it is necessary to consider a continuous approach based on the response to edge cases in actual operation and post-marketing management.

6.2.2. Product Performance

To comprehensively evaluate the product performance of GenAI functions in GenAI-enabled medical devices, it is essential to view the capabilities of GenAI functions by breaking them down into three levels: "general language capability and instruction-following", "basic medical knowledge", and "task-specific performance".

In addition, it should be noted that GenAI outputs vary with various factors. An important perspective in the review based on this variability is the difficulty of determining the validity in a single output. In addition, there are many GenAI-enabled medical device use cases for which no single correct answer (ground truth) exists. Therefore, rather than focusing on matching specific "values", an approach that defines the required "properties" of the target behavior is needed. Specifically, "property-based testing" is useful to provide statistical assurance that a property is preserved by searching an enormous automatically generated test space for inputs and operations that violate the property.

In addition, three approaches are typically used to determine and evaluate appropriateness of outputs generated in the verification process.

- Rule-based evaluation, an automatic evaluation using acceptance criteria derived from experiences (heuristic metrics).

- Assessment by physicians or other experts based on clinical context and safety considerations.
- LLM-as-a-Judge evaluation, in which outputs are evaluated by another language model according to predefined evaluation criteria (e.g., rubrics).

It is important to select or combine these methods depending on whether a ground truth is available in the task and the scale of evaluation necessary (e.g., when a property-based test, addressing a large test space, is adopted, automated evaluation methods, such as rule-based evaluation and LLM-as-a-Judge evaluation, are expected to play a central role in the evaluation process. Expert assessment may be considered for sample-based evaluation and clinical judgment of identified violations).

In product performance evaluations, static benchmark tests that measure the basic capabilities of GenAI models or systems (e.g., JGLUE, Japanese MT-Bench) are also useful in understanding the baseline. However, benchmark performance is not necessarily directly linked to the clinical relevance of the product under application, and there are limitations in interpretation of the test scores. Performance can be overestimated if the test dataset or question format is known or if the training dataset includes the benchmark unintentionally (i.e., data leakage). Therefore, it is appropriate to make decisions based not only on benchmark results but on the overall evaluation (e.g., appropriateness of test datasets, evaluation procedures, evaluation methodology, study protocols) according to the clinical positioning of the product under application.

Based on the above, these review points below describe the specific aspects of GenAI's product evaluation, focusing on property-based testing. However, static benchmark testing is still in the mainstream of GenAI evaluation. When designing static benchmarks, refer to considerations set out in [Appendix 3](#).

(1) General Language Capability and Instruction-Following

General language capability and instruction-following are fundamental capabilities including natural language comprehension and responding in line with instructions, such as constraints given by prompts and output formats. Existing benchmark tests to measure these capabilities, which can be used as a reference for basic capabilities of GenAI models, include the following:

- Benchmark tests on understanding Japanese: JGLUE, etc.
- Benchmark tests on generating Japanese: Rakuda, JapaneseVicunaQA, Japanese MT-Bench, etc.

For example, the properties to be maintained by GenAI models can be identified as "accuracy of text outputs", "suitability to context", "instruction-following", "safety", and "reproducibility and stability" and can be verified in the following manner based on the medical context:

1) Accuracy of Text Outputs

Evaluate the degree to which the information generated in response to the input is consistent with objective facts and possible answers. For tasks without ground truths, consider scoring according to appropriate evaluation metrics (e.g., rubrics).

To verify the accuracy of text outputs using property-based testing, it is effective to verify the semantic relations between inputs and outputs (i.e., metamorphic relations). Examples include verifying whether the output stays consistent when input expressions undergo minor modifications, such as paraphrasing and variations in expressions, and whether the output changes appropriately when input conditions, such as symptoms, are logically added or reversed.

In addition, it is also important to verify noise tolerance. Examples include verifying whether unrelated dummy text or intentionally introduced noise does not induce hallucinations, and whether the input can still be processed in line with the original intentions and meanings of the inputs.

2) Suitability to Context

Evaluate whether the response is based on appropriate information in complex contexts including past conversation history, long texts, and presentation of references by RAG.

To verify suitability to context using property-based testing, it is effective to verify semantic relationships. Examples include verifying whether the consistency of output is ensured when past conversation history and references are paraphrased or summarized, or whether the model correctly captures context changes when patient conditions (e.g., presence of comorbidities or changes in situations) are logically updated during the conversation, and whether the generated output changes appropriately.

In addition, it is also important to verify noise tolerance. Examples include verifying whether the model can maintain a valid output meeting the intended clinical context without overlooking important information, by intentionally replacing the order of references presented by RAG, adding dummy literature information, or by performing operations that are unrelated to past conversation history or intentional noise in the context.

3) Instruction-Following

Evaluate whether the model complies with constraints given by prompts, such as output formats (e.g., JSON formats), character count restrictions, or "do not use a particular vocabulary".

To verify instruction-following using property-based testing, it is effective to verify the semantic relations between inputs and outputs. Examples include verifying whether the generated output consistently meets the same formats and constraints when inputs are changed, such as paraphrasing prompts without changing the logical meaning of constraints or changing the order of multiple instructions.

In addition, it is also important to verify noise tolerance. Examples include verifying whether the specified output format and the system constraints are securely maintained, even if complex requirements are added including long inputs or intentional noise contamination (e.g., typos, omissions, homonymic words).

4) Safety

Evaluate whether the system does not generate harmful information or cause unintended behavior in response to malicious inputs or unexpected requests.

To verify safety using property-based testing, it is effective to verify semantic relations between inputs and outputs. Examples include verifying whether harmful outputs are consistently avoided or interrupted, even when inputs are changed to imply inappropriate requests (e.g., requests for harmful information) or questioned in a fictitious scenario, or whether it will not excessively reject appropriate medical questions that are paraphrased due to false operation of the guardrails.

In addition, it is important to verify noise tolerance. Examples include verifying whether the system can safely reject adversarial prompts (e.g., jailbreaks), such as "ignore previous instructions", and improper requests with large amounts of unrelated phrases and intentional noise, without disturbing functional boundaries and model behavior.

5) Reproducibility and Stability

Evaluate whether the system consistently generates outputs of an acceptable quality level without unexpected errors, considering the probabilistic output variations of GenAI.

To verify reproducibility and stability using property-based testing, it is effective to verify the semantic relations between inputs and outputs. Examples include verifying whether the generated output variation does not deviate from the permissible requirements and maintains an acceptable quality level, not only by repeating the same input multiple times but also by changing input prompts including paraphrasing.

In addition, it is also important to verify noise tolerance. Examples include verifying whether the system maintains the properties without degradation or collapse in model performance, even during extended conversation sessions with long contexts, or when the input is contaminated with irrelevant phrases or intentional noise.

(2) Basic Medical Knowledge

Basic medical knowledge is a fundamental capability specialized in the medical domain to accurately understand and apply disease concepts and medical terminology. Existing benchmark tests to measure these capabilities include IgakuQA using questions from tests such as the National Examination for Medical Practitioners. The percentage of correct responses to these questions is used as a reference metric.

In addition, properties that should be maintained by GenAI-enabled medical devices, such as accuracy of medical information and safety, should be identified. These properties can be verified as follows:

In addition to the accuracy of inferences based on medical consensus and clinical guidelines, evaluate whether the output strictly complies with the requirements concerning contraindications that directly affect the safety of patients (e.g., contraindications to medical procedures and drugs, allergy data). Evaluation is also necessary to ensure that ethically problematic outputs are not generated (e.g., discriminatory expressions and biases).

To verify accuracy of medical information and safety using property-based testing, it is effective to verify the semantic relations between inputs and outputs. Examples include verifying whether consistent medical reasoning is maintained, without recommending contraindicated treatments, even when semantically equivalent questions are expressed with different medical terminology or when the same clinical condition is presented in different scenarios.

In addition, it is also important to verify noise tolerance, such as whether the boundaries of medical information accuracy and safety are securely maintained without inducing hallucinations or harmful outputs, even when unrelated dummy symptoms or complex scenarios with intentional noise were included.

(3) Task-Specific Performance

Task-specific performance refers to the performance of individual clinical tasks claimed by the product under application. This level often does not have generic benchmark tests, so a dedicated test dataset that adequately reflects the diversity of actual clinical settings needs to be constructed (e.g., patient attributes, presence of comorbidities, atypical symptoms).

In addition, the properties to be maintained can be identified as "suitability for complex clinical contexts" and "functional stability in unintended use" and can be verified as follows:

1) Suitability for Complex Clinical Contexts

Evaluate whether outputs conforming to the purpose can be generated without overlooking important clinical information, in clinical setting scenarios that assume multiple background factors rather than a single symptom.

To verify suitability for complex clinical contexts using property-based testing, ensure that critical clinical elements are included in the outputs, even when the order of the patient background, medical history, drug information, or other information is changed, or certain information is paraphrased. It is also effective to verify the semantic relations between the accumulated context and the output, such as whether the generated output is updated in a logical manner according to their importance when clinical information is provided incrementally.

In addition, it is also important to verify noise tolerance, including whether the system can maintain consistent appropriateness without misjudging the priorities of information or compromising the clinical utility of the output, even when the context includes complex scenarios, with unrelated descriptions or peripheral information that does not directly affect clinical decisions.

2) Functional Stability in Unintended Use

Evaluate whether GenAI systems do not present unintended diagnosis or treatment plans in response to user requests beyond the scope of the functions approved for the medical device. This includes ensuring that the implemented guardrail functions properly in response to entries with deviations from the claimed function, due to unintentional functional changes and inputs based on off-label use, inappropriate use, or misuse.

To verify this property using property-based testing, it is effective to verify the semantic relations between inputs and outputs, such as whether the guardrail will operate consistently even when the input is changed to request tasks that are not claimed (e.g., request definitive diagnosis for non-target diseases) using implicit or synonymous medical expressions.

In addition, it is also important to verify noise tolerance, such as whether the boundaries of the claimed function are securely maintained with appropriate fail-safe functions without disturbing the model behavior, even when low-evidence, uncertain information, large volumes of unrelated descriptions, or adversarial prompts with intentional noise are automatically generated and included.

6.2.3. Other Evaluations

In addition to the product performance evaluation described above, it is necessary to evaluate whether risk control measures, such as guardrail functions, implemented as outputs of risk management activities for the product under application are functioning properly. The evaluation would be considered according to the purpose and the intent of risk control measures.

In addition to the above, conformity with the Essential Principles for medical devices must be confirmed, including the software development life cycle (JIS T 2304), the usability engineering process (JIS T 62366-1), and cybersecurity (JIS T 81001-5-1).

6.3. Post-Marketing Measures

6.3.1. Post-Marketing Reporting

It is difficult to cover diverse edge cases that may be encountered in actual operation only with the limited premarket evaluations described above. In addition, considering the dynamic characteristic of GenAI-enabled medical devices (e.g., output variability due to dependence on external reference information), there are fundamental limitations to justify that the results of fixed premarket evaluations continue to ensure appropriate efficacy and safety even after marketing. Moreover, if agentic AI can examine and execute steps autonomously to achieve its objectives, it may respond in methods that the developers had not imagined (including those with ethical and technological risks).

To complement premarket evaluations, applicants would be expected to actively collect information on post-marketing performance and on the occurrence of adverse events and hallucinations, together with periodic reporting of whether these findings raise clinical concerns. If the function is executed in an agentic AI-like manner, problems may arise in the process leading to its execution even when there are no problems in terms of clinical outcomes or adverse events. In such cases, consideration should be given not only to clinically recognizable hazards but also to hazards that are difficult to recognize. Logs of execution by the product under application are also expected to be monitored. It is also important to consider the implementation of systems and functions that can collect such information. These measures need to be considered case by case, according to premarket evaluations and risks of the product under application.

In addition, it is necessary to discuss in advance with the authorities regarding responses to problematic events. In the event of an exceptional incident, it is also necessary to address responses case by case depending on the frequency and impact of problematic events.

6.3.2. Design and Operation based on Human Factors

As described in Section 5.2, the use of GenAI-enabled medical devices requires consideration of limitations in explainability, risks of off-label use, inappropriate use, and misuse, as well as challenges arising from human factors,

such as user literacy and cognitive biases.

However, in traditional medical practices, physicians have been constantly collecting new knowledge from sources, such as literature, judging its appropriateness based on their own expertise, and applying it to clinical practice. Given that physicians critically examine and utilize medical literature that may contain uncertain information, hypotheses that have not yet been validated, or information with low level of evidence, residual incompleteness in the output does not, in itself, preclude approval of the device. Importantly, it is necessary to improve the operational environments in which users can properly judge and interpret the authenticity and appropriateness of information.

The differences between references to GenAI outputs and literature are summarized as follows:

Evaluation Perspective	Conventional Medical Literature and Reports	Output by GenAI	Risks in Output by GenAI
Evidence verifiability	Author, journal, and citation source are clearly indicated, and can be verified.	The source is unclear unless special frameworks, such as RAGs, are used.	Traceability and verifiability have not yet been established.
Accumulation of expert assessments	Reliability through assessments by peer review systems and expert groups, such as academic societies.	Clinical assessment on the output has not been accumulated.	Absence of socially accepted reliability assessment framework.
Knowledge transfer and education	Knowledge on "credible/not credible publication" is shared through education and mentoring.	Educational and empirical basis to determine the appropriateness of outputs has not yet been sufficiently established.	Literacy to identify errors has not yet been established.
Cognitive bias	Physicians are trained to critically review the literature on the assumption that it may contain errors and biases.	Algorithm outputs appear objective, accurate and easy to accept without scrutiny (automation bias).	Risk of overconfidence and dependency resulting from human factors.

In other words, challenges in actual operations include the absence of explainability of GenAI outputs and uniform operational policies to evaluate the reliability of the output, as well as human factors, such the absence of user literacy and automation bias.

Considering the above, it is essential that GenAI-enabled medical devices can be used where users correctly understand the degree of uncertainty and reliability of outputs. Specifically, it is desirable to take the following measures. It is also desirable to incorporate them in design as necessary.

- **Presentation of supporting evidence:** It is desirable to implement a function that informs users whether the generated output has an identifiable source and, if so, what that source is, and whether the output is supported by existing evidence or is an inference extrapolated from existing information.
- **Rejection of off-label use, inappropriate use, or misuse:** For inputs based on information that clearly deviates from the intended use, relates to contraindications, or information with limited evidence, it is desirable that the product under application implements a function with a specification to reject such requests (e.g., guardrails), rather than relying on cautionary statements to prevent such requests.
- **Risk management for unintentional off-label use:** In addition to the intentional inappropriate use described above, there is a risk of unintentional off-label use by the user due to the availability of a wide range of inputs beyond the intended use defined by the marketing authorization holder. In addition to conducting risk assessments based on this point, cautionary statements on proper use must be thoroughly provided at the time of marketing.

- **Educational support to improve user literacy:** It is desirable to provide checklists and training for users to obtain a certain level of literacy for the proper use of GenAI. In particular, there are concerns about "deskilling" in which excessive reliance on GenAI outputs results in degradation of users' critical thinking and professional skills, even if it is more convenient. Therefore, this is an essential measure that should be continuously optimized and maintained in accordance with technological advances and the development of social literacy.

It is also assumed that GenAI-enabled medical devices will be used by the patients themselves by applying the characteristics of GenAI. In such cases, the patient's degree of literacy in medicine and healthcare varies greatly among individuals, making it difficult for users to identify hallucination through their own literacy and to manage risks based on critical scrutiny. Therefore, in addition to establishing user requirements and providing instructions, it is also necessary to consider measures for design specifications of the product under application, such as implementation of a guardrail to prevent inappropriate use or misuse by users and ensure safety through device design (safety design based on the foolproof and failsafe concepts).

7. Conclusion

While GenAI has enormous potential in the healthcare sector, it also contains unprecedented, multi-layered regulatory and technological challenges. The considerations described in these review points are provisional and based on current knowledge. It should be further elaborated by the accumulation of actual product developments and review results.

There are views that the risks of GenAI-enabled products would be deemed unacceptable because of false or erroneous generations due to hallucination. However, healthcare professionals today are already striving to provide the best care through interpretation and judgment based on information supported by varying levels of evidence. A similar approach may be applicable to the outputs of GenAI. However, at this early stage, there is little empirical evidence to determine the extent to which GenAI outputs, which may contain false or erroneous information, can be trusted. Therefore, the impact of hallucination on user judgment or care needs to be monitored carefully. For the time being, it may be necessary to require users to consult with a physician familiar with the subject matter, rather than adopting GenAI's suggestions immediately, if the user cannot appropriately judge the accuracy of GenAI outputs.

In any case, introducing GenAI technologies into healthcare in a positive manner is considered to benefit both healthcare professionals and patients. Developers, regulatory authorities and the medical domain will need to cooperate to promote social implementation of safe and effective GenAI-enabled medical devices.

Appendix 1: Examples of Development Models

Table 1 Examples of Development Models1

Development Model	Features	Areas Typically Subject to Direct Applicant Control
Use of proprietary models (e.g., cloud-based system APIs)	Generate inferences by calling the model provided by third-party developers. Details of model specifications and training data are often undisclosed.	Prompt design, guardrails, input/output design, peripheral components such as RAGs, test dataset/procedures, operation/monitoring (within log availability)
Use of open-weight models (Operated in the applicant's own environment)	Obtain the weights disclosed and generate inferences within the applicant's environment (or in a specific environment). It is easy to lock the inference environment.	Inference environment (hardware/software configuration, version), deployment procedures, access control and logging, implementation reproducibility including model settings (e.g., temperature parameters), security measures.
Additional training (e.g., fine tuning)	Perform additional training on the above models using data for medical applications to adjust their performance and behavior.	Selection and quality of additional training data, training procedures, hyperparameters, retraining and change controls, differential evaluation, version management of training results.
Use of in-house GenAI models	Use GenAI models that are designed and trained by the developer and incorporated into medical devices. It is costly to achieve the necessary performance, but the models can be optimized according to the purpose. Update control is easy to implement.	Model structures, training algorithms, training data selection and pre-processing, training conditions and evaluation methods, model updates, whether retraining is performed, and related procedures and management policies

Appendix 2: Technological Characteristics that may Raise Issues and Suggestions for Development

A. Output Variability

Outputs of GenAI models may vary with various factors and may become non-deterministic. If the behavior of the model varies with inputs having identical meanings, the challenge is to evaluate properties with limited test cases, such as reproducibility of outputs as a medical device, consistency of behavior, and identity of functions. An important perspective in evaluation is to identify the 'variation factors' of these outputs and to organize the dynamics of factor-specific variability (e.g., global or local, continuous or discrete, ability to return to the initial state) along with their degree of control. The degree of control over the dynamics of such variations also depends on how the model is operated, such as the use of proprietary models via API or on-premises systems.

Therefore, based on the dynamics of each variation factor in GenAI models, appropriate risk control needs to be implemented as well as functional designs and other measures after clarifying whether the variability can be locked or can be controlled within predefined ranges and boundaries.

Examples of matters to be considered to analyze variation factors in output are as follows:

(1) Output Variation Arising from User Input

Generated outputs from GenAI models vary with slight differences in expressions of user inputs (prompts), even when they have the same context. This characteristic allows for generation of flexible and expressive content. However, minor changes in inputs, such as abbreviations used in clinical practice, nonstandard medical vocabulary, formats different from those used in development, and incorrect inputs, such as typos and omissions, can lead to variations and decreases in appropriateness of generated inferences, which may confuse the user's decision making. The challenges in evaluation include settings, such as prompts to evaluate models for each task, tolerable range of output variation associated with changes in input, dependency on the user's input skills and addressing reduced reproducibility.

(2) Output Variation Arising from Context

Generated outputs from GenAI models adapt to the context of the interaction logs as well as the prompts entered, causing continuous variation of the output. As a result, although the direction of the output by GenAI may be optimized to the context specific to each user, based on log data or other information recording the actual operation of the product under application, their accuracy may be reduced due to the accumulation of inappropriate context. Such log data and other relevant information may include conversational contexts and log information from current or past use of the product under application.

As the behavior of the model may be dynamically adjusted according to the operation of the product under application, the challenges in evaluation would be the prerequisite context to conduct evaluation and to include matters specified in the approval, such as the initial state or after accumulation of specific interaction logs. In addition, it is important to consider the controllability of the system, such as whether the accumulated context can be reliably reset to its initial state.

(3) Output Variation Arising from Model Settings

In GenAI models, such as LLMs, sentences are generated based on the probabilities of the next occurrence of tokens (the smallest unit in which the model processes text). Rather than constantly selecting the most probable token, they are sampled with constant randomness, resulting in a variation of outputs in response to identical inputs¹. Since unique, expressive outputs are generated from this characteristic, developers are expected to implement it in the product under application. Particularly for the interface of the product under application intended to support treatment for the public, such "human-like" characteristics may contribute to the treatment adherence and efficacy of the product. Therefore, these characteristics are expected to be implemented, even in medical devices.

The challenges in evaluation would be that it is not possible to fully cover the output characteristics of a product, including its variations, when only a single output is evaluated for a given input. It is also critical to identify parameters associated with variations in output and to ensure the stability of system behavior through controllability.

(4) Output Variation Arising from External Systems and Data

For GenAI models that use highly interoperable functions, such as the RAG and agentic AI functions described above, updates to connected databases and external systems will cause corresponding changes in outputs that refer to them (discrete variations). This variation enables tasks, such as diagnostic support, to be based on the latest medical knowledge.

The challenge in evaluation is that the output for the same input varies with evaluation timepoints and reproducibility decreases due to changes in the reference information over time. In addition, when a large amount of external information online is used without restrictions, false/erroneous information or biased information (e.g., health information on Westerners for Japanese patients) may be acquired and incorrect answers may be generated. Therefore, it is necessary to consider them carefully in risk management processes. In addition, controllability over the entire GenAI system, such as locking connected external systems and data to a particular condition, should be considered. This relevant variation may also be a challenge in operation.

(5) Output Variation Arising from Foundation Model Updates

Foundation model developers adjust and update their foundation models on an ongoing basis. Therefore, when a proprietary model is used via cloud API, the output for the same input may change irreversibly with updates in model characteristics due to the relevant model updates. This is positioned as a global variation, associated with changes in the external environment, unlike local variations within the GenAI system from (1) to (4). Therefore, it should be noted that the property and controllability of the variations in (1) to (4) above may change with updates in foundation models.

This can be considered an operational challenge in evaluating whether the medical device remains the same when the GenAI model, a core component of the product under application, is updated.

B. Functional Changes

¹ Such variations in output may be caused not only by intentional introduction of randomness but also by non-deterministic fluctuations in computational processing, such as those at the hardware level.

The variation in output described in section A is not merely a variation in textual output; it can also be a factor that causes "functional changes" in the medical device. The variation factors listed in A are not limited to variations in output quality but may extend to changes in the function of the medical device.

For example, adjustments to system prompts at each installation site in accordance with clinical protocols (discrete functional changes) and personalization based on the user's past interaction logs (continuous functional changes) may be positioned as functional optimizations that take advantage of GenAI's capabilities to generate content depending on the context.

However, if this adaptability is not controlled, models may deviate from approved claimed functions by adapting excessively to unconstrained inputs and accumulated context. Specifically, when the user enters questions about target diseases or clinical intentions that have not been verified in their prompts, there is a risk that GenAI models will generate plausible answers, resulting in functioning as unverified diagnostic and treatment support.

Therefore, for functional changes in these GenAI models, the dynamics of the change need to be organized along with their controllability, like the variations of the outputs. Perspectives including the tolerance of system prompt adjustment for users, and how to monitor and control continuous functional changes based on use logs would be necessary.

In addition, it is important to design reasonable test data to ensure the efficacy and safety of the claimed function, implement guardrails to prevent off-label and inappropriate use beyond the claimed function, and evaluate the system considering human factors.

C. Accuracy of Outputs and Suitability to Context

GenAI-enabled medical device outputs may have a significant impact on patients' lives and health, depending on the positioning and the intended use of the product under application. Therefore, it is essential to ensure that the generated information has the following aspects: (1) "suitability to context", whether it is an appropriate response for a particular context, and (2) "accuracy of outputs", whether the generated information is consistent with medical facts and evidence, as well as with context including complex clinical context or user intentions.

In addition to tasks for which a correct answer can be defined and predictive accuracy can be quantitatively evaluated, such as "detecting findings on images," GenAI-enabled medical devices are also expected to generate appropriate responses in contexts where no single correct answer exists, such as appropriate word selection according to the literacy, psychological status, social background, etc. of the user.

To evaluate "accuracy of outputs" and "suitability to context", it is important to note that the usefulness of GenAI content is strongly dependent on the user's interpretation (interpretation dependency). In addition, it is necessary to assess significant risks, such as false information, harmful information, and guidance that can distort users' interpretations, as well as the effect of data biases inherent in GenAI models.

(1) Interpretation Dependency of Generated Outputs

GenAI content contributes to clinical outcomes, such as improved diagnostic accuracy and treatment outcome, through proper interpretation of the GenAI content generated in response to user inquiries. In other words, unlike conventional Discriminative AI in which outputs are expressed by labels, the user must interpret the content generated

by GenAI in natural language texts or other formats in terms of its appropriateness to the medical accuracy and context and select useful information.

A challenge in evaluation is how to evaluate GenAI content, in natural language texts or in other formats, in terms of reproducibility and objectivity. In addition, considering the characteristic that the authenticity of the generated output is left to the user's interpretation, the design and operation of the GenAI-enabled medical device should consider the risks inherent to GenAI described below.

(2) False, Harmful Information and Guidance

The output generated by GenAI may contain "false information" or "erroneous information" due to hallucination. In addition, the generated content may have a direct negative effect, such as "harmful information" or "inappropriate guidance".

A challenge in evaluation is how to objectively measure the output tendency of false, erroneous or harmful information for each specific intended use scenario. Another challenge is the tolerance of outputs that may lead to inappropriate guidance, and how to establish and evaluate guardrails to prevent them.

(3) Data Bias

Data bias is not limited to GenAI but is a characteristic of any product developed using data-driven machine learning technologies. However, products implementing GenAI models that use extremely large datasets are more difficult to identify and control data biases compared to conventional machine learning, because they use larger and diverse non-standardized data and are trained widely without supervised labels. Implicit values and cultural assumptions may be included unintentionally and may affect the efficacy and safety of medical devices (e.g., recommended actions may assume that users live in urban areas because urban data is overrepresented in the training data, the generated information is inappropriate for young patients due to the large amount of data for elderly patients).

A challenge in evaluation is that inputs and outputs in the use case assumed in GenAI-enabled medical devices are extremely diverse, making it difficult to cover, in advance, how data bias may be actualized. Therefore, risk management specific to GenAI-enabled medical devices would be necessary.

Appendix 3: Positioning and Limitations of Benchmark Testing

A. General Description of Benchmark Testing

Benchmark testing is an evaluation method that quantifies model performance by scoring its response to a predetermined question or dataset based on specific criteria. In GenAI (particularly LLM), it is widely used to understand aspects of comprehension, reasoning, and generation in a comparable manner.

The main significance of benchmark testing lies in (1) relative comparison between models, (2) understanding the direction of improvements, and (3) improving explainability through standardization of evaluation procedures.

However, since benchmark testing measures performance in a particular question format and evaluation condition, it does not directly demonstrate efficacy and safety considering the clinical positioning of the product under application. Therefore, benchmark results should be interpreted in conjunction with evaluations considering the intended use and use condition of the product under application (e.g., test datasets, procedures, study protocols).

B. Limitations of Benchmark Testing (Notes for Evaluation)

At a minimum, the following considerations and limitations should be taken into account when planning an evaluation using benchmark tests:

- Deviation from clinical utility: Even if a high score is obtained in benchmark tests, the performance may not be sufficient to make decisions necessary in clinical practice (e.g., incomplete information, differences in prerequisites, contraindications/exceptions, accountability). For example, some studies have quantitatively examined the consistency between medical question-answering (QA) benchmarks and performance and efficacy in clinical practice, suggesting that such deviations may occur (Siun Kim, Hyung-Jin Yoon. Proceedings of the 24th Workshop on Biomedical Language Processing. 2025). It has also been reported that structured testing for triage recommendations may reveal performance variations and blind spots depending on the test conditions (A. Ramaswamy, et al. Nature Medicine. 2026).
- Dataset contamination/overestimation by known problems: If the test dataset or similar data is included in the training data, models may answer correctly by "memory" rather than "capability", and its performance may be overestimated. Especially in open benchmarks, it is difficult to eliminate the risk of unintentional training (contamination risk).
- Leaderboard overfitting: When the benchmark becomes the optimization target, improvement in a particular benchmark score may not be generalized to other tasks or to actual operating conditions. There are limitations to discussing performance on a single score.
- Limitations of evaluation indicators and scoring methods (reproducibility and objectivity): Correct answers often cannot be uniquely defined for tasks in GenAI, and the results may vary depending on rating indicators and scoring rules. Manual evaluation can lead to variation between evaluators. Even when using automatic evaluation (e.g., LLM-as-a-Judge), attention should be paid to matters including the consistency of the judgment, bias, settings of the LLM for evaluation, and dependence on prompts.
- Missing important safety failures: Even when the average accuracy rate is at a certain level, rare but significant errors may be overlooked (e.g., misleading guidance on contraindications, errors in emergency determination). It is difficult to guarantee safety by using benchmarks alone.

C. **Importance of Appropriate Benchmark Design for the Target Task and Design Examples**

The appropriate design of benchmarks differs depending on the target task (e.g., diagnostic support, treatment support, triage, documentation support), intended users and use conditions. Based on the clinical positioning of the product under application, it is essential to define the capability to evaluate before designing benchmarks.

- **Clarification of tasks to evaluate:** Organize the value provided by the product under application (e.g., differential diagnosis support, contraindication checking, generation of explanation, and summarization) and define input conditions as well as the format, granularity, and prerequisites of the expected outputs.
- **Clarification of critical requirements:** For aspects where errors can lead to significant results (critical requirements), such as contraindications, safety, and emergency determination, consider not only the average score but also the design to capture failures in individual cases (e.g., focused sampling, thresholds, error analysis).
- **Definition of correct answers and design of acceptance criteria:** Predefine rationale (e.g., guidelines, package inserts, standardized textbooks) and scope of correct answers (e.g., tolerable expressions, synonymous wording, conditional answers). State them in study protocols.
- **Management of evaluation conditions and assurance of reproducibility:** Ensure reproducibility of evaluation results by defining settings such as prompts (including content and granularity of expression), tools to be used (e.g., RAG, external DB), and the range of presented information.
- **Multifaceted evaluation:** Evaluate from multiple aspects according to the task, rather than rely on single scores (e.g., accuracy, completeness, consistency, avoidance of contraindications, appropriateness of justifications).

D. **Basic Perspectives in Designing Benchmarks**

In designing benchmarks, the technological issues inherent in the factuality evaluation of GenAI content must be considered in addition to the properties of the evaluation samples (e.g., coverage and representation).

- **Coverage and representation of evaluation samples:** Structure evaluation samples to represent intended use conditions, including variations such as target disease/symptoms, severity, comorbidities, age groups, etc. Include rare but important cases intentionally (e.g., contraindications, emergency responses).
- **Medical accuracy and relevance:** Clarify the basis for the correct answer and enable evaluation of the scope of medically appropriate expressions and conditions (e.g., prerequisites, exceptions, contraindications). Incorporate expert reviews as necessary.
- **Technological aspects to evaluate the factuality of GenAI content:** Factuality is evaluated based on the following aspects: (1) scope of the evidence to be referred to, (2) consistency with the evidence, and (3) handling uncertainty (e.g., avoid affirmative expressions, conditioning, and clarification of limitations). Even when the output is plausible, the evidence may be insufficient. Therefore, the presence and the appropriateness of the evidence may be included as evaluation items (e.g., consistency in citations).
- **Reproducibility and auditability of scores:** For manual evaluation, clarify evaluator requirements and acceptance criteria. When conducting automatic evaluation (e.g., LLM-as-a-Judge), fix conditions, such as the

LLM used for evaluation, prompts, acceptance criteria, sampling conditions, and cross-check with manual evaluation as necessary.

It is crucial to make comprehensive determinations not only with benchmark results, but also with evaluations considering the clinical positioning of the product under application.

Therefore, while benchmark testing provides valuable information, it should not be overemphasized as a surrogate for clinical utility or safety.

References

- ⁱ “Good machine learning practice for medical device development: Guiding principles” (IMDRF/AIML WG/N88 FINAL:2025, January 29, 2025)
- ⁱⁱ "Report on AI-based Software as a Medical Device (SaMD)" (Subcommittee on Software as a Medical Device Utilizing AI and Machine Learning, PMDA Science Board, dated August 28, 2023)
- ⁱⁱⁱ "Guidance for the Evaluation of Emerging Technology Medical Devices Attachment 4 "Guidance for Evaluation of Artificial Intelligence-Assisted Medical Imaging Systems for Clinical Diagnosis"" (PSEHB/MDED Notification No. 0523-2, Director of Medical Device Evaluation Division, Pharmaceutical Safety and Environmental Health Bureau, Ministry of Health, Labour and Welfare, dated May 23, 2019)
- ^{iv} "Regulatory Science on AI-based Medical Devices and Systems" (Subcommittee on AI, PMDA Science Board, dated December 27, 2017)
- ^v Moor M, Banerjee O, Abad ZSH, Krumholz HM, Leskovec J, Topol EJ, Rajpurkar P. Foundation models for generalist medical artificial intelligence. *Nature*. 2023 Apr 12. doi: 10.1038/s41586-023-05881-4.