Overview of the Risk-Based Credibility Assessment Framework for AI Models Proposed by the FDA

Japan Pharmaceutical Manufacturers Association,
Drug Evaluation Committee
Electronic Data Management Committee Task Force 4
March 30, 2026

[Disclaimer]

The information contained in this document is based on information available at the time of publication. The Japan Pharmaceutical Manufacturers Association (JPMA) assumes no responsibility for any damages or losses resulting from the use of this

< Revision History >

Version	Date	Reason for Revision
1.0	March 30, 2026	Initial Release

Table 1: Terminologies

Term	Definition
AI Model	Mathematical algorithms that learn patterns (relationships between inputs and outputs) from data to optimize parameters (weights). These algorithms may be positioned as an "AI Subsystem" or "Component" incorporated into a broader IT system (Core System).
AI System	A machine-based system that operates with varying levels of autonomy. For explicit or implicit objectives, it infers from the inputs it receives to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.
Algorithm	A mathematical procedure or formula for learning patterns (features) from data to solve a problem. It refers to the "learning method" itself used to generate an AI model (e.g., Random Forest, Neural Networks).
Architecture	The internal structure and design specification of an AI model. It refers to the structural framework that defines how a model performs computation and inference, such as the number of layers and connection methods in a neural network, or the depth of a decision tree.
Calibration	In the context of AI, the process of making fine adjustments to the output of a trained model aimed at improving accuracy and/or repeatability. This differs from "calibration" in the GxP

Term	Definition
	context, which refers to the comparison and adjustment of measuring instruments against a reference standard.
Credibility	The degree to which an AI model's performance can be trusted to be adequate for its intended purpose within a specific context of use (COU). Unlike "validation" in conventional system development, this is a concept proposed by FDA that accounts for the development methods and uncertainties inherent to AI models.
Credibility Assessment	A set of activities conducted to establish the credibility of an AI model within a specific context of use (COU). By systematically evaluating the model development process, performance, quantification of uncertainty, and impact on risk, it determines whether the model is fit for use for the COU (broad sense). In this document, the term is also used in a narrower sense to refer to the testing activity that confirms the selected model meets its intended performance, corresponding to testing (or UAT) in the traditional CSV framework.
Context of Use (COU)	A specification of the specific role and scope of an AI model, detailing "what" it is used for and "how" it is applied. It serves as the basis for determining the model's risk and the required level of Credibility.
Data Drift	A phenomenon in which the statistical characteristics of operational input data diverge from those of training data over time or due to environmental changes, or the relationship between inputs and outputs changes (concept drift). As a primary cause of AI model performance degradation often accompanied by Model Drift, continuous monitoring is required for early detection.
Data Integrity	A framework of requirements to ensure the trustworthiness of data collected, analyzed, stored, and reported in GxP activities. Based on the ALCOA++ principles, it encompasses a broad range of requirements such as data handling, identification and access management, audit trails, electronic signatures, and security.

Term	Definition
Data Leakage	A situation in which test data, which should be reserved exclusively for evaluation, is inadvertently used in model development or tuning. Both the FDA AI Guidance and Annex 22 emphasize Data Independence among training, tuning, and test datasets, and Data Leakage is regarded as a serious risk representing a failure to maintain the independence of test data.
Decision Consequence (DC)	The severity of consequences on subsequent regulatory decision-making, patient safety, or product quality in the event of an incorrect AI model output. It is one of the two components that constitute model risk.
Explainability	The degree to which the output and reasoning of an AI model, particularly a black-box model, can be understood and interpreted by humans. When supporting critical decision-making under GxP, the ability to explain how a conclusion was reached is essential for ensuring credibility.
Feature	A data attribute used as input when training a model. In tabular data, it corresponds to a column (e.g., height, weight, sex). The values contained in each column are referred to as feature values.
Fit for Use	The state in which data is relevant and reliable (i.e., accurate, complete, and traceable) for its intended purpose. As AI model performance is highly dependent on the quality of training data, data being fit for use is a fundamental prerequisite. The term also refers to the state in which a model's performance is deemed adequate and appropriate for its intended COU.
Grid Search	One of the methods used to find the optimal combination of hyperparameters for an AI model, primarily a machine learning model. It is the process of exhaustively testing all combinations of predefined hyperparameter candidates and their ranges, and selecting the combination that yields the best performance.
Human in the Loop (HITL)	A process in which humans, such as subject matter experts, review and verify the output of an AI system and are involved in final decision-making. It is an important control measure for correcting AI errors and mitigating risks such as black-box opacity.

Term	Definition
Intended Use	The purpose defining "what a system or process is used for." It is a fundamental principle of validation in GAMP® 5 Second Edition. In the context of AI systems, it is required to be further specified as a "Question of Interest (QoI)" and a "Context of Use (COU)."
Life Cycle Maintenance	A set of planned activities to manage intentional or unintentional changes, ensuring that an AI model remains fit for use within its COU. It encompasses the monitoring and maintenance of model performance and suitability throughout the lifecycle.
Machine Learning (ML)	A subset of AI techniques in which algorithms are trained to improve performance on a task by optimizing model parameters based on data through computational processes, rather than through explicit programming.
Model Influence (MI)	The degree to which the output (evidence) of an AI model influences a regulatory decision relative to other evidence considered in the overall decision-making process. It is one of the two components that constitute model risk.
Model Risk	The potential for an AI model's output to lead to incorrect decisions, resulting in adverse outcomes for patient safety or product quality. FDA proposes that it be assessed through a combination of Model Influence (MI) and Decision Consequence (DC).
Multiple Models	Multiple model candidates developed in parallel during the AI model development phase (exploratory development) to achieve optimal performance. They are generated through different combinations of algorithms, architectures, and hyperparameters, and serve as the basis for comparative evaluation.
Overfitting	A state in which an AI model is excessively fitted to development data, resulting in degraded predictive performance (generalization performance) on unseen data not used during training. It is one of the primary risks that undermine model credibility.

Term	Definition
Question of Interest (QoI)	A specifically and clearly defined problem or question that an AI model is intended to address or answer. It serves as the starting point for defining the scope and goals of AI model evaluation.
Risk-Based Approach	An approach to determining the rigor and scope of lifecycle activities based on risks to patient safety, product quality, and data integrity. It is a fundamental principle of GAMP® 5 and a central concept in the assurance of AI.
Selected Model	The model selected from multiple candidates, having been judged through evaluation using tuning data to have the best performance meeting the COU, and designated as the subject of final testing (Credibility assessment). The "specification" of this model is ultimately adopted as meeting the COU.
Static / Dynamic / Adaptive Model	**Static Model:** A model whose parameters are fixed (frozen) after training and do not change during operation. Re-training is performed only as part of a version update through a planned change management process. **Dynamic Model:** A model that has an "internal state" that changes over time, used in time-series or sequential decision-making contexts. Parameters may be fixed; a model is considered dynamic if its internal state is updated with each inference. **Adaptive Model:** A model whose parameters are automatically updated during operation in response to new data or environmental changes. Annex 22 is scoped to static and deterministic AI models only. Dynamic/adaptive models and models with probabilistic outputs are outside the document's scope. For these model types, Annex 22 explicitly states they **should not be used** in critical GMP applications that directly affect patient safety, product quality, or data integrity. As a result, AI models available for critical GMP use are effectively limited to static and deterministic models.
Test	The activity of objectively evaluating the performance (generalization performance) of a trained AI model using

Term	Definition
	independent test data. It corresponds to activities such as "Step 5" in the FDA Credibility assessment framework.
Test Data	A completely independent dataset used to objectively evaluate the final performance of an AI model after development and tuning are complete (also referred to as "hold-out data").
Training Data	A dataset used to build and train an AI model. As it determines the foundation of model performance, the quality and representativeness of the data are of critical importance.
Tuning Data / Validation Data	A dataset used to evaluate trained models and select (explore) optimal hyperparameters and model architectures. What FDA AI Guidance refers to as tuning data is called a Validation Dataset by Annex 22 and GAMP® 5. In general machine learning literature, "validation set" is used synonymously to refer to data used for model selection and hyperparameter tuning. In this document, following FDA terminology, this dataset is consistently referred to as "tuning data" to avoid confusion with "validation" in the GxP context, which refers to system qualification and verification activities.
Validation	A series of processes to verify and document that a computerized system as a whole meets predetermined requirements and specifications (Intended Use). In this document, the term is used as an overarching concept that encompasses not only the performance evaluation of an AI model itself (Credibility assessment), but also the conformance confirmation of the entire system incorporating the AI model (core system and AI subsystem).
Verification	The activity of confirming that what has been developed (in this case, an AI model or system component) has been built correctly in accordance with its specifications and requirements. In this document, the term is used as an umbrella term encompassing all AI model performance confirmation activities (evaluation, testing, and verification); however, in practical contexts, it is treated as synonymous with "Test."

Note: The definitions in this glossary prioritize facilitating understanding of international guidance documents such as the FDA AI Guidance, Annex 22, and GAMP® 5, and therefore adopt

basic and limited explanations of mathematical and technical AI concepts. Definitions set forth in each company's internal procedures shall take precedence.

Table 2: List of Abbreviations

Abbreviations	Definition
AI	Artificial Intelligence
CAPA	Corrective and Preventive Action
COU	Context of Use
CSA	Computer Software Assurance
CSV	Computerized System Validation
DC	Decision Consequence
DWH	Data Warehouse
EC	European Commission
EMA	European Medicines Agency
FDA	Food and Drug Administration
GAMP®	Good Automated Manufacturing Practice
GCP	Good Clinical Practice
GMP	Good Manufacturing Practice
GVP	Good Vigilance Practice
GxP	Good x Practice
HITL	Human in the Loop
LLM	Large Language Model
MI	Model Influence
ML	Machine Learning
PoC	Proof of Concept
QMS	Quality Management System
QoI	Question of Interest
SaMD	Software as a Medical Device
SME	Subject Matter Expert

Table 3: List of Trademarks

Trademark or Registered Trademark	Trademark Holder
GAMP®	International Society for Pharmaceutical Engineering (ISPE)
Docker®	Docker, Inc.
TensorFlow®	Google LLC

Note: Other company names and product names mentioned in this document may be trademarks or registered trademarks of their respective owners.

Table of Contents

1. Introduction

In recent years, AI technology has advanced rapidly, significantly expanding its scope of application within the pharmaceutical industry. From drug discovery research to clinical development, manufacturing, and post-marketing safety surveillance, AI has the potential to deliver innovative value throughout the entire drug product life cycle. Concurrently, driven by diversifying and increasingly sophisticated global medical needs, as well as the necessity to respond to unforeseen crises such as pandemics, the demand for more rapid and efficient drug development is growing. In this context, utilizing AI is no longer merely an option; it is becoming an essential undertaking for pharmaceutical companies to fulfill their social responsibilities and improve patients' quality of life (QOL).

However, the implementation of AI in the pharmaceutical industry presents inherent challenges. In drug development and manufacturing, ensuring patient safety, product quality, and data integrity are absolute prerequisites, guaranteed by strict regulatory requirements such as GCP, GMP, and GVP. When implementing AI in these GxP environments, specific methodologies for assuring quality and demonstrating credibility have not yet been fully established. This is largely due to AI's unique properties, such as data-driven learning processes, complex behaviors, and "black box" characteristics, which differ significantly from those of traditional computerized systems.

Against this background, the U.S. Food and Drug Administration (FDA) issued a draft guidance titled "Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products" ("FDA AI Guidance") in January 2025, proposing a risk-based Credibility Assessment Framework for AI models. Additionally, the European Commission (EC) published a draft of "EU GMP Annex 22: Artificial Intelligence" ("Annex 22") in July 2025, presenting more specific technical requirements. These regulatory trends require pharmaceutical companies to respond to a new paradigm of AI quality assurance, creating an urgent need to promote understanding and implement practical responses across the industry.

This document aims to provide basic concepts and practical guidelines for AI validation to a wide range of stakeholders, including AI practitioners involved in the implementation, development, and operation of AI in pharmaceutical companies, as well as those who are not experts in AI technology. While centering on the 7-Step Risk-Based Credibility Assessment Framework proposed by the FDA, this document seeks consistency with international standards such as Annex 22, the "AI Guidelines for Business" issued by Japan's Ministry of Economy, Trade and Industry (METI), and "ISPE GAMP® 5: A Risk-Based Approach to Compliant GxP Computerized Systems (Second Edition)"

("GAMP 5 2nd Edition"). It explains the comprehensive overview of AI validation, integrating the FDA's framework into the GAMP 5 life cycle phases.

The preparation of this document was guided by the reality that "AI technology is evolving day by day, and regulatory requirements are changing in real-time." Both the FDA AI Guidance and Annex 22 referenced herein are in the draft stage, and future revisions or the issuance of additional guidance are expected. Therefore, this document is not intended to provide a definitive "answer," but rather to serve as a "thinking framework" to guide decision-making when pharmaceutical companies establish their approaches to AI quality assurance according to their own situations. Critical thinking remains paramount; it is essential to continuously address questions such as: "Why use this AI?", "What are the risks if the AI produces incorrect results?", and "How will those risks be managed?".

Task Force 4 of the Electronic Data Management Committee of the Japan Pharmaceutical Manufacturers Association is committed to continuously monitoring changes in regulatory trends and sharing knowledge and discussions within the industry. It is sincerely hoped that this commentary and the collection of case studies will help promote understanding among practitioners in pharmaceutical companies and assist in practical AI quality assurance activities. Furthermore, by being flexibly utilized within each company's QMS, this document aims to contribute to accelerating innovation in drug development and ultimately providing better medical care to patients.

March 2026
Japan Pharmaceutical Manufacturers Association, Drug Evaluation Committee
Electronic Data Management Committee Task Force 4

2. Scope and Purpose of This Document

2.1 Intended Readers

The purpose of this document is to summarize recent advancements in AI technology and accompanying regulatory trends, and to provide practical guidelines for the utilization of AI systems in pharmaceutical companies. This document is intended for "AI Users" within pharmaceutical companies who utilize AI systems, and "AI Validation Practitioners" who are involved in the implementation, development, and validation (CSV) activities of AI. For the former, this document provides a general explanation to foster a basic understanding of AI systems. For the latter, it serves as a practical guide supporting CSV activities for AI system implementation and operation. Throughout this document, the term "practitioners" refers collectively to AI Validation Practitioners in this latter group.

While the FDA AI Guidance sets out "Credibility Assessment" requirements for regulatory submissions, this document applies those requirements more broadly, as practical guidance for anyone working with AI technology across the pharmaceutical product life cycle.

2.2 AI Models Covered by This Document

While this document aligns with the FDA AI Guidance framework for "Regulatory Decision Support Tools," it also encourages extending these concepts to other use cases at the sponsor's discretion. A clear distinction should be made between "explicit regulatory expectations under the FDA AI Guidance" and "practical extensions and interpretations presented in this document"; the latter is not intended to supersede regulatory requirements.

Generally, AI technologies are classified as shown in Figure 1.

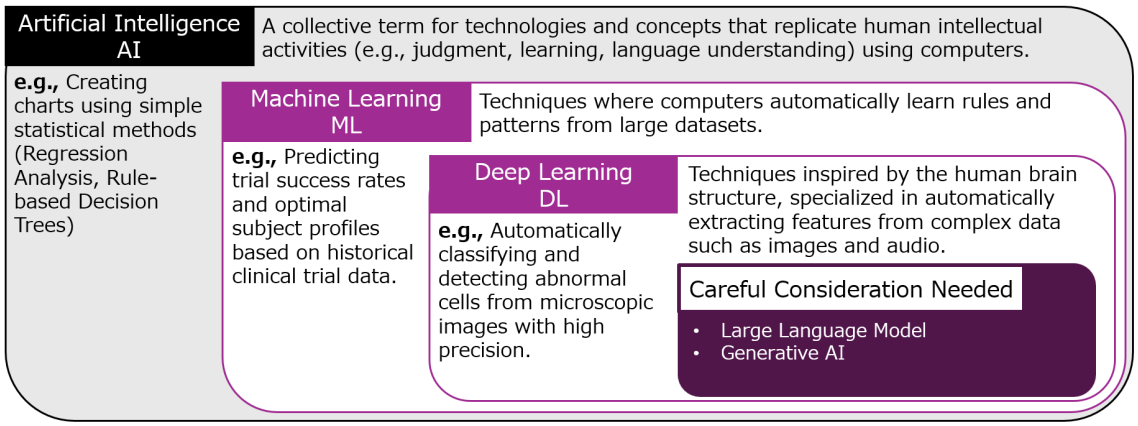

Figure 1: AI Categorization

The FDA AI Guidance primarily targets "AI models" that learn from data. This document also targets the following:

- Machine Learning (ML) models
- Deep Learning (as a subset of ML)
- Large Language Models (LLM)
  - While Large Language Models (LLMs) are included in ML and Deep Learning, the FDA AI Guidance primarily envisions traditional machine learning models (e.g., prediction, classification, detection) and does not include specific requirements specialized for LLMs or Generative AI. Therefore, when using LLMs to support regulatory decision-making, sponsors should carefully consider explainability, reproducibility, and version control within the general framework of the FDA AI Guidance, while acknowledging their inherent "black box" nature. Early consultation with regulatory authorities is highly recommended.
  - It is also important to note that Annex 22 explicitly states that models with probabilistic outputs, such as LLMs and Generative AI, **should not be used** in critical GMP applications (see Section 6.).

In this document, AI models that "learn parameters from data" are collectively referred to as "AI Systems" or "AI Models." To aid understanding, these AI models are classified by mechanism and output type as follows:

- **Classification by Mechanism:**
  - **Static Model:** A model whose learned parameters are fixed during operation and are not changed even when new data is acquired. Re-training is performed only as part of a version update through a planned change management process.
  - **Dynamic Model:** A model that has an internal state which changes over time, used in time-series or sequential decision-making contexts. Even if parameters are fixed, a model is considered dynamic if its internal state is updated with each inference.
  - **Adaptive Model:** A model whose parameters are automatically updated during operation in response to new data or environmental changes. This can be combined with either static or dynamic models.
- **Classification by Output Type:**
  - **Deterministic Model:** A model that always returns the same output for the same input and internal state. Most AI model inference processes are implemented in this sense.
  - **Probabilistic / Stochastic Model:** A model that explicitly represents prediction of variability as probabilities or distributions, where outputs may vary stochastically even

for the same input and state. This includes cases where the output itself is a probability or distribution. For example, a probabilistic model defines a probability distribution over outputs based on the input data, and derives the output by sampling from that distribution.

These mechanism and output classifications are independent of each other; various combinations exist, such as "static and deterministic," "dynamic but deterministic," "static but probabilistic," and "adaptive and stochastic."

Regardless of these combinations, this document broadly targets "AI models that learn parameters from data." However, considering practical utility, static and deterministic models are the primary focus. Dynamic and adaptive AI models are discussed in detail in Section 6. in the comparison between the FDA AI Guidance and Annex 22.

Conversely, the following technologies are excluded from the scope of this document:

- Rule-based expert systems
- Simple decision logic consisting solely of if-then rules
- Simple statistical modeling that has been widely used in the GxP context (e.g., interpretable linear regression models, determinations based on predefined thresholds or scores)
  - **Note:** While simple statistical methods are excluded from detailed explanation, practitioners are encouraged to apply the principles of this document to high-risk Contexts of Use (COU) and consider Credibility Assessment in line with the FDA AI Guidance as necessary.

3. Basic Principles and Overview of the Credibility Assessment Framework for AI Systems

This section outlines the overall picture and basic concepts regarding the necessity of activities within the Credibility Assessment Framework. This explanation is intended for personnel engaging in Credibility Assessment Framework activities for AI systems for the first time. This explanation complies with the FDA AI Guidance and references Annex 22 and the METI "AI Guidelines for Business" ("AI Business Guidelines"). Furthermore, these basic concepts are explained considering consistency with the concepts of ISPE's GAMP 5 2nd Edition. This section serves as the foundation for understanding the specific processes detailed in Section 4. and beyond.

3.1 AI Models Within the Scope of the FDA AI Guidance

First, the scope of the FDA AI Guidance is identified. Figure 2 illustrates our interpretation for identifying target AI models subject to the FDA AI Guidance.

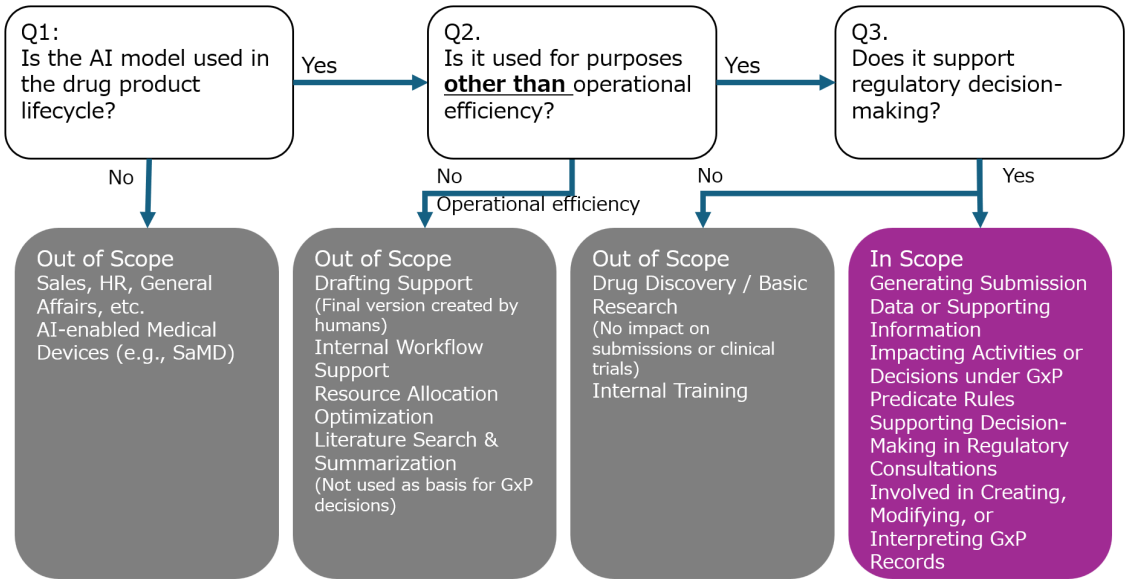

Figure 2: Decision Flow for Applicability of FDA AI Guidance

The FDA AI Guidance applies to AI models whose outputs support regulatory decisions by the FDA and/or pharmaceutical companies across the drug product life cycle. Conversely, AI models that remain at the drug discovery or basic research stage, or that are used purely for operational efficiency without affecting regulatory decisions or the reliability of GxP records, are excluded from the scope. To make that determination, the process is structured into a flow consisting of the following three questions, in accordance with the FDA AI Guidance.

Q1: Is it used in the drug product life cycle?

Sponsors first determine whether the data handled by the AI model is used for the drug product life cycle, such as drug development, manufacturing, or post-marketing safety measures. Business efficiency tools unrelated to pharmaceuticals, such as sales or HR, are considered out of scope at this stage. Additionally, medical devices such as Software as a Medical Device ("SaMD") are excluded because they fall under the framework of medical device-related guidance, not the FDA AI Guidance.

Q2: Is it used for purposes other than operational efficiency?

If the AI model is used solely for operational efficiency purposes, such as drafting GxP documents, supporting internal workflows, or resource allocation, it is considered out of scope.

Q3: Does it support regulatory decision-making?

Sponsors determine whether the AI model supports regulatory decisions by the FDA and/or pharmaceutical companies across the drug product life cycle, covering non-clinical studies, clinical studies, manufacturing, and post-marketing safety management.

If the answer is "No," it is out of scope. Examples include drug discovery, basic research, internal training other than GxP, and literature searches. If the answer is "Yes," the AI model is classified as a "Regulatory Decision Support Tool" subject to risk-based Credibility Assessment and lifecycle management.

The level of rigor, however, should reflect the model's potential impact on patient safety and product quality. For models with lower direct risk, such as those supporting Pharmaceutical Quality System (PQS) operations, the rigor of Credibility Assessment and related activities may be adjusted based on the risk evaluation described in Section 5.2 .

For these "Regulatory Decision Support Tools," Section 3.2 explains the Credibility Assessment Framework. This explanation is based on the basic principles of the life cycle approach, risk-based approach, and ensuring data integrity.

3.2 The Need for Credibility Assessment in AI Systems

When implementing AI systems in drug development or manufacturing, the output results may directly affect patient safety, product quality, and the reliability of results. Therefore, pharmaceutical companies need to ensure, with documented evidence, that the AI system functions safely and effectively as intended. This is the objective of the activities within the Credibility Assessment Framework for AI systems.

The background necessitating Credibility Assessment Framework activities specific to AI systems, rather than traditional validation represented by CSV, stems from risks unique to AI models and fundamental differences from traditional software.

1) Risks Unique to AI Models

AI models contain risks not found in traditional software, such as:

- **Bias Risk:** If the training data contains bias, the AI model's outputs will reflect that bias, potentially producing inaccurate results for specific populations.

- **Lack of Transparency Risk:** Humans may not fully understand the rationale for why the AI model reached a specific conclusion, leading to a "black box" state. When unexpected errors occur, investigating the cause and implementing countermeasures becomes difficult.
- **Ambiguity of Accuracy Risk:** It becomes difficult to interpret, explain, or quantify the correctness of the results output by the AI model.
- **Performance Degradation Risk:** After starting operation, there is a risk that "data drift" (covariate shift) occurs where the statistical distribution of actual data (e.g., manufacturing equipment, patient background) changes, or "concept drift" occurs where the relationship between input and output changes. In this document, these are collectively referred to as "Data Drift," but practitioners should distinguish between them when analyzing causes or considering countermeasures.

2) Differences from Traditional Software

Traditional software is a deterministic system that follows clear instructions created by humans and always returns the same result. On the other hand, AI models learn rules and patterns from large amounts of data to construct decision rules. Therefore, the behavior of AI models has the following characteristics:

- **Strong Dependence on Data:** The performance of an AI model depends on the quality and quantity of data used for learning, not on pre-defined logic. If learned with inappropriate data, the AI model's output will also be inappropriate.
- **Complexity and Difficulty in Interpretation:** Depending on the learning method and internal parameters, the AI model's algorithm becomes complex, and it may be difficult for humans to predict or grasp all of it. Therefore, not only the results but also the validity of the learning process and robustness against patterns of data encountered for the first time in the real environment (which have not been learned) are required.

This "dependence on data and self-learning" explains why quality cannot be assured by traditional deterministic testing alone. For this reason, while the FDA AI Guidance uses "Validation" and "Reliability" for systems and data, it introduces a new concept called "Credibility" for AI models. This suggests that the FDA requires framework activities to continuously ensure a credible state, taking into account the specific characteristics of AI models.

3.3 AI System Life Cycle

To assess AI model credibility in accordance with the FDA AI Guidance, it is helpful to first understand the overall life cycle of the system in which the AI model operates. Understanding where

the AI model fits within the computerized system life cycle described in GAMP 5 Second Edition enables more effective Credibility Assessment activities. Therefore, this section first explains the system architecture and the overall picture of the life cycle.

Credibility Assessment requires viewing the full series of activities, from AI system conception through development, operation, and retirement, as a managed "life cycle" in which AI model performance is continuously evaluated. To appropriately assess the life cycle, AI Users and practitioners will need to understand the composition of the "AI system," which typically consists of three distinct elements: the Core System, the AI Subsystem, and the AI Model.

3.3.1 System Architecture: Relationship between Core System and AI Model

First, understanding the architecture of the target AI system is essential. AI system configurations are diverse, and the following patterns can be considered:

- Architecture where the AI model is integrated into the Core System
- Architecture where the AI model cooperates as an independent subsystem
- Architecture utilizing cloud-based AI API services

Based on GAMP 5 2nd Edition, this document presents an example configuration where the Core System (GxP System) and the AI Model are clearly separated and interact via an AI Subsystem. Actual system architectures should be selected according to each company's IT environment, risk assessment, and technical constraints.

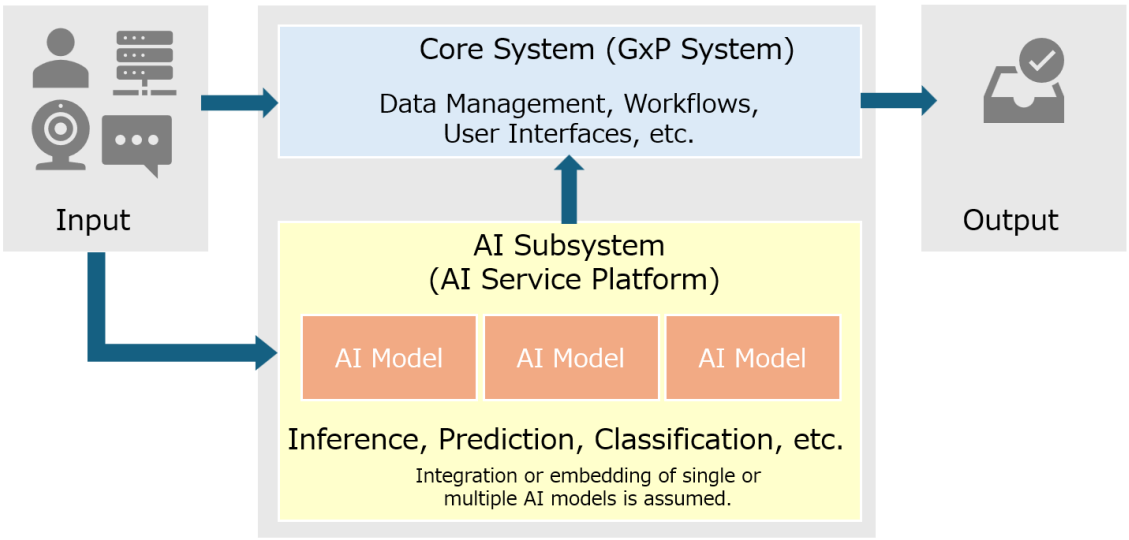

Figure 3: Example Architecture of Systems Incorporating AI Models

- **Core System (GxP System):** The Core System is the main application or platform operated directly by users in a GxP environment, such as Manufacturing Execution Systems (MES) or clinical data analysis platforms. It encompasses functions including data management, business workflows, user interfaces, and audit trails. In the context of traditional CSV, the Core System constitutes the primary "computerized system."
- **AI Subsystem (AI Service Platform):** The AI Subsystem is a platform hosting one or more AI models and facilitating their integration with the Core System. While AI models may be directly embedded in the Core System, this document adopts an architecture in which the Core System and AI Model interact through a dedicated AI Subsystem as the baseline configuration.
- **AI Model:** The AI Model is a component specialized for executing specific intelligent tasks such as prediction, classification, and detection. It is typically invoked by the Core System, receiving data from internal and external sources, performing processing, and returning results. Examples include product defect/non-defect judgments and future risk scores.

The relationship between the Core System and the AI Model can be understood by viewing the AI Model as a "plug-in" or "service" integrated into the Core System. The Core System and the AI Model serve distinct roles: the Core System provides a validated, stable business foundation, while the AI Model contributes specific advanced functions. The data flow to the AI Model may follow various patterns, such as via the Core System or directly from external sources, but in all cases, it is essential that the system functions as a coherent whole.

Sponsors will need to verify that the Core System correctly exchanges data with the AI Model, and that the system reliably satisfies its original "Intended Use."

3.3.2 The Three Phases of the AI Model Life Cycle

The life cycle and management approach differ fundamentally between the Core System and the AI Model.

- **Core System Life Cycle:** The Core System follows a traditional software development life cycle. Validation follows standard CSV practices, focusing on areas such as business and functional requirements, data integrity, security, and audit trails. Validation also confirms connectivity with the AI Subsystem, the output of AI Model results, and that the overall process enables end-users to make appropriate regulatory decisions based on a proper understanding of the AI Model's output.
- **AI Subsystem Life Cycle:** As the AI Subsystem serves as the underlying platform, validation focuses on verifying that the AI Model operates correctly, confirming connectivity with the AI Model, and ensuring configuration management.

- **AI Model Life Cycle:** The AI Model life cycle is fundamentally dynamic and iterative. To address performance degradation during operation (Model Drift) or the availability of new training data, retraining, updating, or replacement of the model may occur depending on the usage environment. Credibility Assessment focuses on areas such as model performance metrics and data quality.

Therefore, in the validation of the entire AI system, sponsors will need to ensure not only that the Core System and the AI Model function correctly on their own, but also that "robustness can be maintained in an integrated state." This assumes that validation activities following the V-model life cycle, as presented in GAMP 5 Second Edition, are applied to both the Core System and the AI Subsystem at a comparable level of rigor. However, AI model development is fundamentally different from traditional software development. For details on AI model-specific development methods, please refer to Section 4.2 .

The following describes the activities and data flow for each phase, for configurations where an AI Subsystem is incorporated into a GxP system, or where a GxP system outputs results through an AI Subsystem in coordination with an AI Model (see Figure 4).

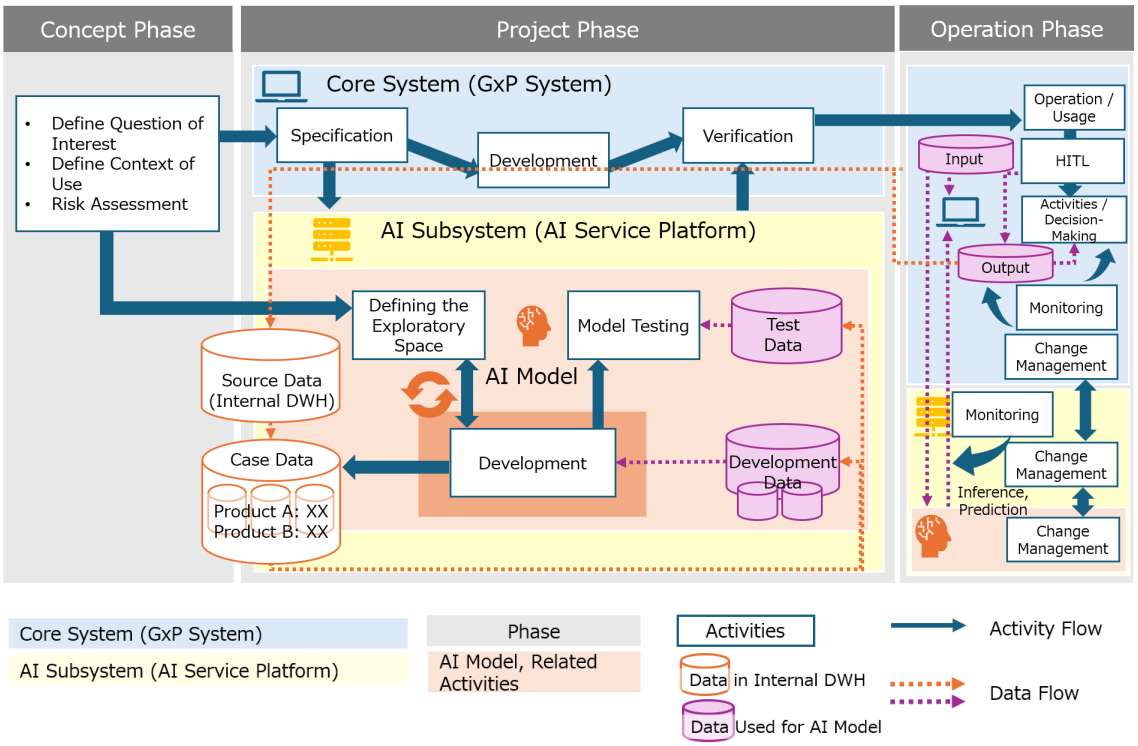

Figure 4: Lifecycle of Core Systems and AI Models

Note: The acquisition and management pathways for inputs and outputs vary depending on the configuration of the overall AI system, including the core system. Examples are provided below:

- Databases for inputs and outputs are integrated into the core system.
- Inputs and outputs are maintained in independent databases separate from the core system.
- Inputs are obtained from an internal Data Warehouse (DWH) where various data are collected and integrated. Outputs are also stored in the internal DWH.

The activities and data flow for each phase are explained below, assuming cases where an AI Subsystem is incorporated into a GxP system, or where a GxP system outputs results in coordination with an AI Model through an AI Subsystem. The general overview of each phase is described below.

1) Concept Phase

This phase marks the starting point of AI validation. The intended use of the AI system, that is, "what the AI system is to be used for," is defined, along with the underlying challenges, operating environment, and constraints. As AI system outputs cannot always be trusted blindly, regulatory decisions and subsequent activities should proceed not solely on the basis of AI system outputs, but also in light of expert judgment and relevant information. In this phase, candidate AI models are explored and developed through an iterative process, with comparison against existing models and processes. Feasibility is then evaluated, informed by a Proof of Concept (PoC) that provides an initial understanding of the scope and performance of the AI model.

Furthermore, the impact of AI Model outputs on decision-making related to patient safety, product efficacy, and quality is analyzed within the context of the intended business operations, and potential risks are assessed. The definitions and evaluations established in this phase determine the quality and direction of all subsequent activities.

2) Project Phase

This is the stage where specific development and verification of the AI model are conducted to achieve the Intended Use defined in the Concept Phase. First, data for training the AI model is prepared. Next, based on the PoC results, the search range for model architecture and hyperparameters is determined. Rather than fixing detailed implementation specifications in advance, models are developed exploratorily to meet performance metrics, and the resulting optimal model becomes the specification. The most critical activity is the performance evaluation of the completed model. Using data not used for training, practitioners verify that the model meets the intended performance.

This verification activity corresponds to testing (or UAT) in the traditional CSV framework and is documented in the "Credibility Assessment Plan" required by the FDA AI Guidance. In this document, "Credibility Assessment" is used as the equivalent of testing (or UAT) in the traditional CSV framework. However, unlike conventional software testing, Credibility Assessment encompasses a broader set of evaluation criteria specific to AI models, including the adequacy of QoI and COU definitions, data quality, training and tuning processes, and model performance metrics. This mapping is adopted to help practitioners integrate Credibility Assessment into existing GxP validation workflows.

A characteristic activity in AI model development is the strict management of data used for training and testing. Ideally, this data should represent the actual environment and target population, meaning it has high similarity to the data used in actual operation and minimal bias. Additionally, data management procedures are implemented to ensure explainability regarding how the prepared data was processed and utilized. Processing activities include labeling for training, such as adding metadata and identifying ground truth features. Utilization activities include splitting data for purposes such as training and verification, and the actual use of specific data for training and verifying specific models and versions.

3) Operation Phase

This is the stage where the core system, including the verified AI model, is utilized in the actual business environment. Where a process incorporating expert judgment and relevant information alongside AI model outputs was established in the Concept Phase, an environment ensuring the reliable execution of that process is put in place. A process where humans are actively involved in final decision-making is primarily referred to as Human-In-The-Loop (HITL).

Quality assurance of the AI system does not end at deployment. Even after actual operation begins, sponsors will need to continuously monitor whether AI system performance is maintained. If initial performance degrades due to environmental changes or trend shifts over time (Model Drift), the cause must be investigated, and maintenance activities such as model retraining or modification must be performed. Through these activities, the credibility of the AI system is maintained throughout its entire life cycle.

4. Overview of FDA Steps and Machine Learning Model Development Methods

4.1 Overview of the FDA AI Guidance Steps

In this Section, the comprehensive activities of the Credibility Assessment Framework, ranging from the conception to the operation of the AI system, are explained in 12 Steps (Step 0 to Step 11). This framework integrates the "Risk-Based Credibility Assessment Framework" proposed in the FDA AI Guidance with the life cycle approach of GAMP 5 2nd Edition.

While the FDA AI Guidance's 7-Step Risk-Based Credibility Assessment Framework primarily focuses on the Credibility Assessment of AI models, implementing and operating AI models in an actual GxP environment requires a series of essential activities. These include PoC (Step 0) for verifying feasibility, integration of the AI model with the Core System (Step 8-9), validation activities leading up to the release of the Core System (Step 10), and operation start and life cycle maintenance (Step 11). In this document, the FDA AI Guidance's 7-Step framework is extended to present a systematic framework of 12 Steps that integrates these activities. Furthermore, the framework from the FDA AI Guidance is combined with "Figure 2: Example of AI Learning and Utilization Flow" from the Attachment (Supplementary Material) of the METI AI Guidelines for Business to clarify the relationship with data flow. Figure 5 details Figure 4, illustrating the 12-Step activities, data flow, and their relationship with documentation.

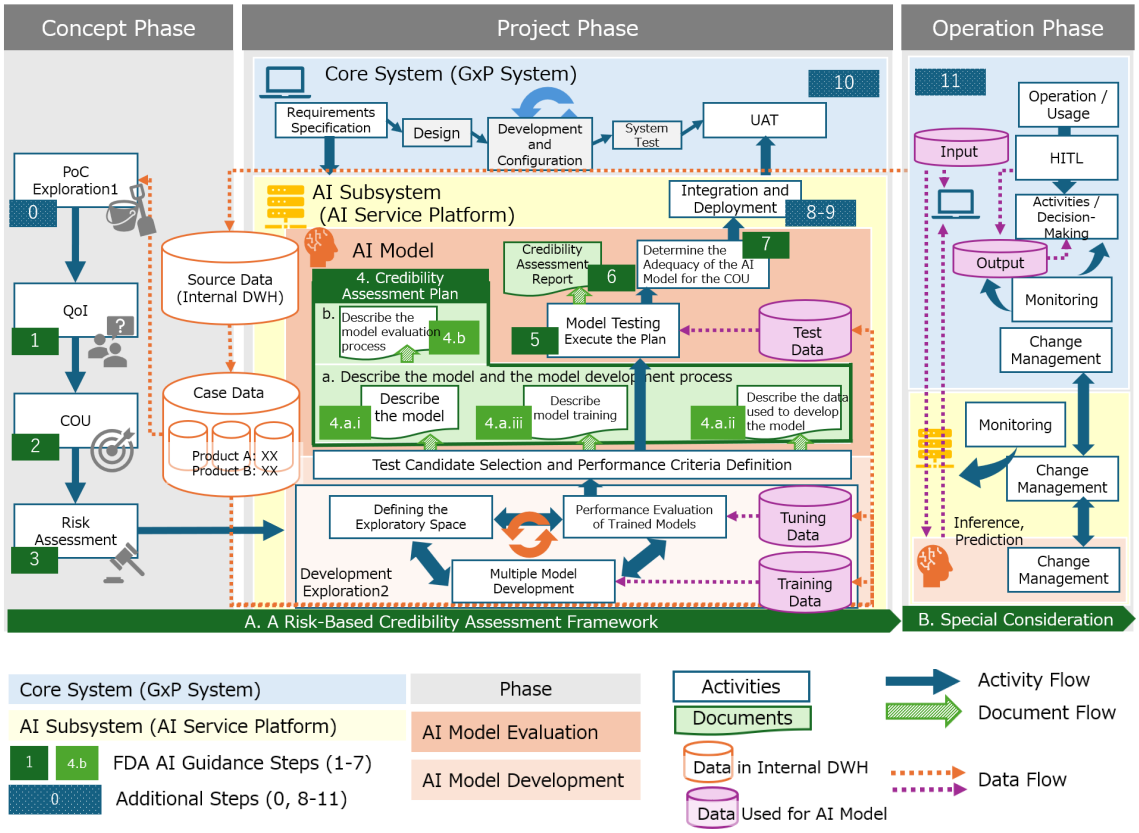

Figure 5: Activities in the AI Model Lifecycle

Note: Figure 5 extends the 7 Steps of the FDA AI Guidance into 12 Steps.

The green boxes in the figure indicate each Step and section of the FDA AI Guidance. The FDA AI Guidance presents the activities of the Concept Phase and Project Phase up to release as "A. A Risk-Based Credibility Assessment Framework" in 7 Steps (Step 1-7), and presents the activities of the Operation Phase in "B. Special Consideration: Life Cycle Maintenance of the Credibility of AI Model Outputs in Certain Contexts of Use."

In this document, in addition to the FDA's 7 Steps, a total of 12 Steps are configured based on the life cycle concepts of GAMP 5 2nd Edition. These include PoC Implementation (Step 0), Construction and Verification of the AI Subsystem and Integration with the Core System (Step 8-9), Validation Activities up to Core System Release (Step 10), and Operation and Maintenance (Step 11). Step 11 is positioned as an operation phase encompassing "Hypercare" immediately after release and subsequent routine operation.

Steps uniquely added in this document are indicated by blue (dotted) boxes. Furthermore, documents and content required by the FDA are indicated by green document shapes.

Section 5. provides detailed guidance on each Step.

4.2 Target AI Models in this Document

The purpose of this document is to present practical methods for assuring the Credibility of AI models that support regulatory decision-making, based on the FDA AI Guidance. Therefore, it assumes that pharmaceutical companies will develop small to medium-scale AI models using internal operational data. Additionally, since high explainability is required for AI models supporting regulatory decision-making, this document envisions binary classification models that offer a balance between explainability and predictive performance in practice, such as XGBoost, Random Forest, and Neural Networks ("NN").

In addition, as shown in Figure 5, AI model development is "exploratory" and "iterative." Therefore, the concept of exploration includes not only the selection of model architecture (algorithm) and adjustment of hyperparameters (degree and method of training), but also data preprocessing and data engineering, such as the selection or transformation of data features (database column names), handling of missing values, and scaling.

Furthermore, this document assumes a development flow where, for a specified model architecture, multiple candidate models are developed from training and tuning data using techniques such as grid search within a predefined hyperparameter range. These candidates are based on predefined performance metrics (e.g., AUC, Sensitivity, F1 Score), and the best model among them is selected as the final test candidate. This approach is also applicable to other hyperparameter optimization methods other than grid search (e.g., random search, Bayesian optimization, etc.)

Regarding terminology, the structure of the algorithm or neural network itself is referred to as the "Model Architecture (Design)," while the mathematical formula with specific parameters learned from data is referred to as the "Model (Finished Product)." In grid search, which is one of the exploration methods, it is assumed that multiple models (candidates) are developed with different hyperparameter settings for the same model architecture, and the best model is selected from them.

4.3 AI Model Development Methods

Among the steps in the FDA AI guidance, the process from Step 4 (Planning Credibility Assessment) to Step 5 (Testing) poses a particular challenge for many CSV practitioners. This is because, unlike the linear system development procedures of the traditional V-model commonly used for computerized systems, ML model development is based on an exploratory and iterative process. Therefore, if practitioners lack sufficient understanding of this iterative development method, developing AI models that comply with the FDA AI Guidance will be challenging. To provide the foundational knowledge necessary to understand the testing in Steps 4 and 5 of the FDA AI Guidance framework, this section organizes the general development process of AI models leading up to these steps and illustrates their relationship in Figure 6.

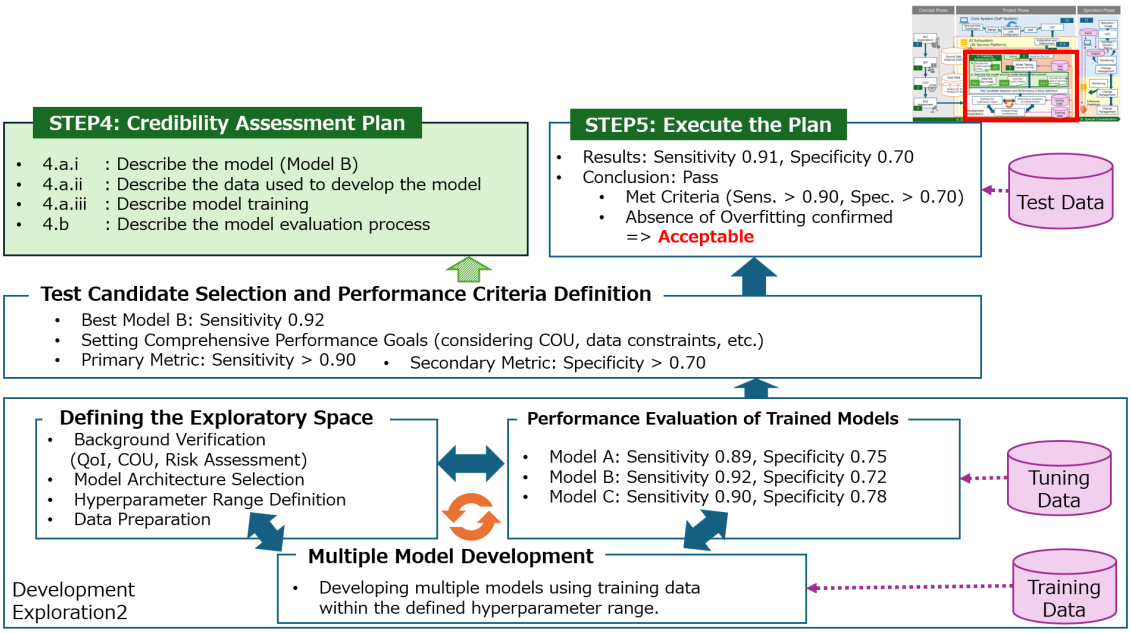

Figure 6: Relationship between Exploratory Development and Testing (Steps 4, 5)

Note: This figure details the activities ranging from "Exploratory Development (Development Exploration 2)" to Step 5, as presented in Figure 5.

Exploratory model development involves creating multiple models within a defined search range, evaluating the performance of each trained model using tuning data, and selecting the best one. If performance satisfying the COU is not reached, the process starts again from the definition of the search scope, including the selection of model architecture and review of hyperparameters. This work is repeated iteratively to select the model that satisfies the COU as a test candidate. Then, the overview of the selected model (4.a.i), the overview of the data used for development (4.a.ii), the overview of model training/tuning (4.a.iii), and the model evaluation process (4.b) are described in the Credibility Assessment Plan, and the test is executed (Step 5).

The important principles for understanding such an ML model development process are as follows:

- **Exploratory Development:**
  - An iterative process of constructing and comparing multiple models in parallel.
  - Developing multiple models using training data and comparing their performance using tuning data.
  - Exhaustively trying all predefined combinations of hyperparameter candidates and their ranges to develop multiple models (assuming grid search).
  - Scientific trial and error to select the best candidate.

- **Data Splitting and Usage:**
  - **Training Data:** Used to develop multiple models.
  - **Tuning Data:** Used to compare trained models and select the best model.
  - **Test Data:** Used to verify the generalization capability (check for overfitting) of the selected model.
- **Differences from Traditional System Development Methods:**
  - The AI model development process is fundamentally different from the development process using the traditional V-model. Specifically, while the traditional method is a sequential and deterministic process of "Pre-determine specifications > Development > Verification," AI model development is an exploratory and iterative process of "Define search scope > Develop/Evaluate multiple models > Select the best." For example, if the "Learning Rate" is set to 3 types (0.001, 0.01, 0.1) and "Regularization" is also set to 3 types (0.1, 1.0, 10.0) as the hyperparameter range, 9 models (3 x 3) are created. Thus, the process of developing multiple models while varying development conditions is why it is called "exploratory." Examples of activities in exploratory development are illustrated in Figure 7.

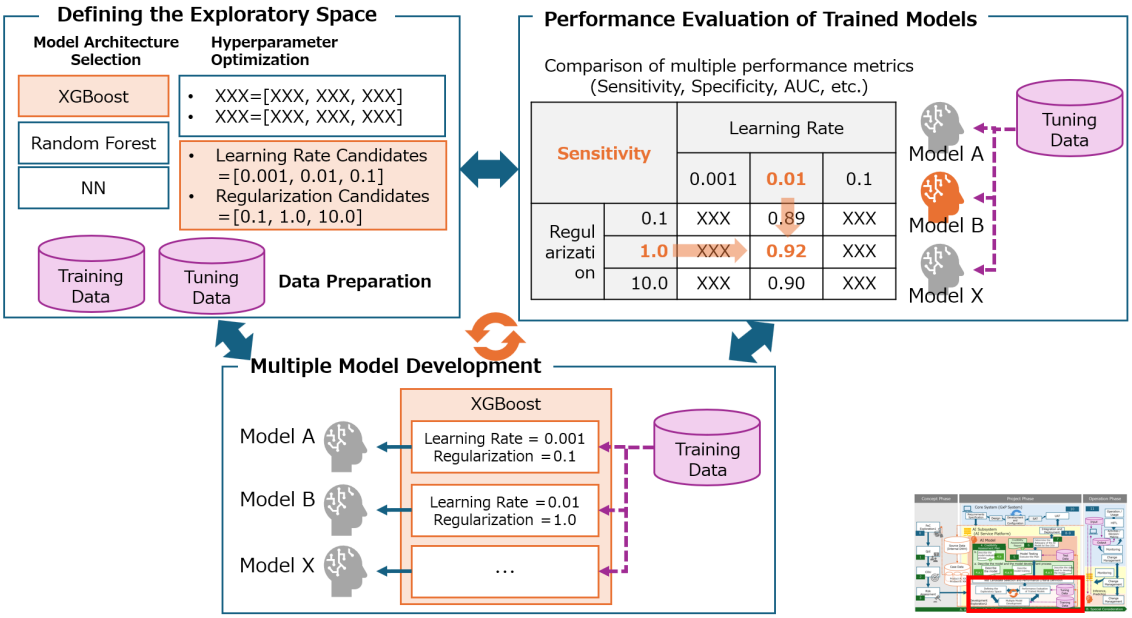

Figure 7: Example of Activities in Exploratory Development

Note: This figure details the activities of "Exploratory Development (Development/Exploration 2)" as presented in Figure 5.

- **Background for Developing Multiple Models:**

  The performance of an AI model is determined by complex interactions between the model architecture, hyperparameters, and data characteristics, which cannot be predicted in advance. Therefore, an exploratory approach of "actually developing and comparing multiple candidate models" is scientifically rational. Considering this exploratory nature, a process of "Definition of Search Scope > Development and Evaluation of Multiple Models > Selection > Testing" is applied.

- **Performance Metrics and Acceptance Criteria:**

  Since the selected model has a potential risk of overfitting, practitioners will need to confirm its generalization capability using independent data. The performance evaluation metrics and acceptance criteria for the test (verification/evaluation) of the selected model should be established scientifically and objectively, comprehensively considering the following elements based on the COU:

  a) **Alignment with COU Requirements:** e.g., Does it meet the direction of emphasizing Sensitivity?
  b) **Data Constraints:** e.g., Consideration of sample size (N=1,200) and positive rate (15%).
  c) **Achievability:** e.g., The best candidate achieved a Sensitivity of 0.92 on tuning data; therefore, 0.90 is a realistic candidate for the performance metric.
  d) **Clinical Utility:** e.g., Do experts judge that Sensitivity = 0.90 is clinically useful enough?
  e) **Statistical Reliability:** e.g., Can the goal be met even when considering confidence intervals?
  f) **Comparison with Current Process:** Is performance equivalent to or better than the current process?

  Taking all of the above into account, sponsors establish performance targets such as: "Primary Metric: Sensitivity ≥ 0.90, Secondary Metric: Specificity ≥ 0.70."

- **Integrated Management with the Core System:**

  In the validation of the entire AI system, in addition to the AI model functioning correctly on its own, sponsors will need to ensure that the model's Output is correctly displayed in the Core System and that "robustness is maintained in an integrated state." In other words, validation activities throughout the life cycle are required for both the Core System and the AI Model respectively.

5. Credibility Assessment Framework Activities: Step-by-Step Guide

In this section, the activities for each Step are described in detail. This section follows a two-part structure: an **"Overview"** based on the FDA AI Guidance, and a **"Discussion"** referencing Annex 22, relevant regulations, the METI AI Guidelines for Business, and GAMP 5 2nd Edition.

Also, as shown in Section 4.2 , the general development procedure for AI models differs from the order of Steps and documentation structure presented in the FDA AI Guidance. Where necessary, practitioners are encouraged to refer to the process in Section 4.2 and associate the order of activities in each Step with the structure of documentation.

5.0 Step 0: Proof of Concept

When deploying an AI system into a GxP environment, before initiating the activities of Steps 1-7 based on the FDA AI Guidance, it is necessary to first determine the applicability of the Credibility Assessment Framework to the AI system in question and establish the framework for the entire project. This is the objective of Step 0. Although not explicitly defined in the FDA AI Guidance, this Step is included in this document as a preparatory stage for AI system implementation, aligning with the concept of the "Concept Phase" in the GAMP 5 2nd Edition system life cycle approach.

[Discussion]

Step 0 is not merely a procedural formality, but a strategic decision-making phase that determines the success or failure of the AI project. The implementation of an appropriate Step 0 can significantly reduce rework in subsequent Steps.

In Step 0, the following major activities are performed:

1) Determination of Applicability

In accordance with the FDA AI Guidance, determine the applicability of this guidance to the AI model in question. The main criteria are as follows. Referring to Figure 2, the guidance applies to AI models that meet all the following criteria:

Q1. Is it used in the drug product life cycle?

Q2. Is it used for purposes other than operational efficiency?

Q3. Does it support regulatory decision-making?

Not all uses of AI models fall within the scope of the FDA AI Guidance. For example, internal operational efficiency tools are out of scope. Conversely, AI models that control data or processes affecting clinical study results, support release decisions in manufacturing, or contribute to safety

assessment and reporting are likely to be in scope. Furthermore, even for AI models within the scope, such as those supporting the operation of the Pharmaceutical Quality System (PQS), if they do not directly affect patient safety or product quality, the rigor of activities should be determined based on the evaluation results of Step 3 (Risk Assessment) described in Section 5.3 .

In principle, the use of AI in drug discovery and basic research stages is out of scope. However, if results from these stages are included in regulatory submission materials or serve as the basis for clinical trial design, they fall within the scope. Examples of potential in-scope applications include:

- AI used for the analysis of non-clinical data included in IND applications
- Prediction models serving as the rationale for dose setting in first-in-human studies
- Biomarker analysis affecting endpoint setting in clinical trials

If determination is difficult, consultation with the Quality Assurance or Regulatory Affairs department is recommended, as well as consultation with regulatory authorities if necessary. Incorrect applicability determination carries the risk of lacking necessary quality assurance or causing delays in AI model development due to excessive activities.

2) PoC Planning and Results

Identify operational effectiveness, the feasibility of AI model construction, and potential issues. Through a PoC that constructs a candidate AI model, sponsors assess the potential to proceed to full-scale development. This PoC typically covers the following:

- Business process overview and background information
- Expected performance and effects (including comparison with existing processes and metrics)
- AI model architecture candidates
- Data to be used (data sources, features, data pathways for inputs and outputs, etc.)
- Supplier requirements (scope of outsourcing, roles, etc.)

Since data is a critical element for AI model development, data management practices should be implemented during the PoC phase. Figure 5 illustrates a process of development and performance evaluation using internal data stored in a Data Warehouse (DWH) available for secondary use. Additionally, sponsors should clarify the scope of outsourcing to suppliers and required capabilities during PoC implementation.

3) Go/No-Go Decision for Project Initiation (Step 1)

Evaluate the validity of proceeding to Step 1 based on the PoC results.

- Technical feasibility
- Adequacy of assumed development data (training and tuning data)
- AI model performance evaluation

Furthermore, identify requirements, recommendations, issues, and risks revealed during the PoC.

5.1 Step 1: Define the Question of Interest

[Overview]

The FDA AI Guidance begins with Step 1, "Define the Question of Interest" ("QoI"). The objective of Step 1 is to define, in clear and unambiguous language, "specifically what judgment the AI model is expected to make" or "what issue is intended to be resolved."

The QoI serves as the foundation for all subsequent steps (COU definition, risk assessment, model design, performance evaluation, etc.). An ambiguous QoI can lead to inappropriate AI models, inadequate test data, and ultimately, an AI system with low credibility.

The following examples illustrate QoI in the GCP and GMP fields.

Example in the GCP Field (Investigational Drug A)

- Background:
  - Investigational Drug A carries a risk of serious adverse reactions. Therefore, conventionally, all subjects were hospitalized for 24 hours after dosing for monitoring to ensure safety. However, this procedure imposed a significant burden on both medical institutions and subjects. Historical data indicates that some subjects are at low risk for adverse reactions.
- Activities and Decision-Making Using the AI System:
  - To accurately distinguish between high-risk and low-risk subjects in advance using an AI system.
  - To conduct efficient and safe clinical trials by implementing outpatient monitoring after dosing instead of hospitalization for low-risk subjects.
- QoI:
  - "Which participants can be considered low risk and do not need inpatient monitoring after dosing?"

Example in the GMP Field (Injectable Drug B)

- Background:
  - In the manufacturing line for injectable drugs, the fill volume of drug solution into vials is visually inspected. Visual inspection imposes a heavy load on inspectors and carries the risk of human error.
- Activities and Decision-Making Using the AI System:
  - To automatically inspect the fill volume of all products at high speed and with high precision using AI-equipped cameras, and to reliably detect defective products (excess or deficiency in fill volume) at 100%.
- QoI:
  - "Do vials of Drug B meet established fill volume specifications?"

To answer the QoI, sponsors will need to establish a process that allows for correction even if the AI results are incorrect. Regulatory decisions or subsequent activities should not rely solely on AI inference or prediction results; instead, they should be determined by combining AI outputs with various related information and evidentiary sources ("Evidentiary Sources"). Evidentiary Sources include, for example, various data generated from in vitro testing, in vivo testing, clinical trials, or manufacturing process validation studies, as well as human judgment results or results from other processes.

The identification of Evidentiary Sources evolves across Steps. In Step 1, potential candidates are identified when defining the QoI. In Step 2, they are specified when defining the COU. In Step 3, they become key inputs to risk assessment, as they determine how much the AI model's influence on decisions can be corrected or mitigated.

[Discussion]

If the QoI is ambiguous, the AI model design, training data selection, and evaluation criteria may become inconsistent, making it difficult to build a credible AI model. The schematic diagram of the process for clarifying the QoI is shown in Figure 8.

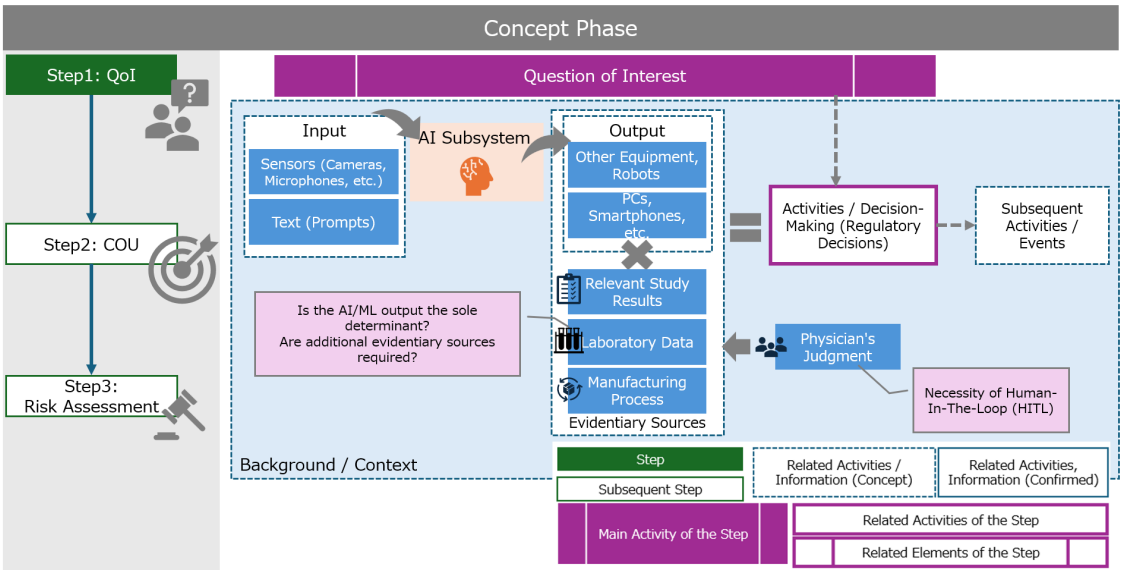

Figure 8: Step 1 Define the Question of Interest

The Step of defining the QoI is similar to "identifying business issues," but it differs in that it utilizes an AI model as the solution. Detailing the business process utilizing the AI system and deeply examining its usage method are considered activities for Step 2 (defining the COU), rather than Step 1. Defining the QoI serves to identify elements required for detailed examination in the COU.

As shown in Figure 8, the QoI in the GCP case was defined as "Which participants can be considered low risk and do not need inpatient monitoring after dosing?". In practice, the final decision on the necessity of hospitalization may also incorporate a physician's judgment. In the GMP case, the QoI was defined as "Do vials of Drug B meet established fill volume specifications?" Here too, it is assumed that the release decision is made in conjunction with visual inspection results.

Evidentiary Sources for answering the QoI are identified in Step 1 and specified in Step 2. Examples of Evidentiary Sources are shown below:

- Existing clinical data and literature information
- Other diagnostic and laboratory results
- Judgment by experts (physicians, pharmacists, inspectors, etc.)
- Outputs from other AI models or systems

When defining the QoI, it is recommended that business departments, data scientists, and quality assurance departments collaborate through iterative discussions to satisfy these criteria.

5.2 Step 2: Define the Context of Use for the AI Model

[Overview]

Step 2 involves defining the COU of the AI model, specifically the "role and scope of the AI." To answer the QoI defined in Step 1, this step concretizes "what Output the AI model produces and how it is used" together with Evidentiary Sources. When defining the COU, the following two elements must be clarified:

- **Role of the AI Model:**
  - The specific tasks the AI model perform such as risk prediction, defect detection, or image classification.
    e.g., Risk prediction, defect detection, image classification, etc.
- **Scope of the AI Model:**
  - The conditions of use for the AI Output, and the presence or absence of other Evidentiary Sources.
    e.g., Whether the decision is determined solely by the AI Output, or whether it is one piece of reference information assisting human judgment.

Table 4 provides specific examples of COU.

Table 4: Examples of COU

Examples	QoI	Role	Scope
GCP: Investigational Drug A	Which subjects can be identified as "Low Risk", allowing for the waiver of hospitalization after administration of Investigational Drug A?	Classify (stratify) the risk of adverse reactions associated with Investigational Drug A, particularly life-threatening risks, into "High Risk" or "Low Risk" based on subject data (patient background, clinical lab values, etc.).	The AI system's Output serves as the sole basis for determining whether a subject requires inpatient monitoring or outpatient monitoring.
GMP: (Injectable Drug B)	Do vials of Drug B meet the established fill volume specifications?	Analyze images of all vials flowing through the manufacturing line to detect vials where the fill volume	The AI system serves as a screening tool for 100% inspection. However, the final shipment decision for

Examples	QoI	Role	Scope
		deviates from the standard.	the product Lot is based on independent verification performed by personnel on a sampling basis, as per conventional methods.

[Discussion]

The purpose of identifying the COU is to define the role and scope of the AI model. The background information (Input, Output, and other Evidentiary Sources) identified in Step 1 is concretized, and the COU is defined by decomposing it into two elements: role and scope. The concept of Step 2 is shown in Figure 9.

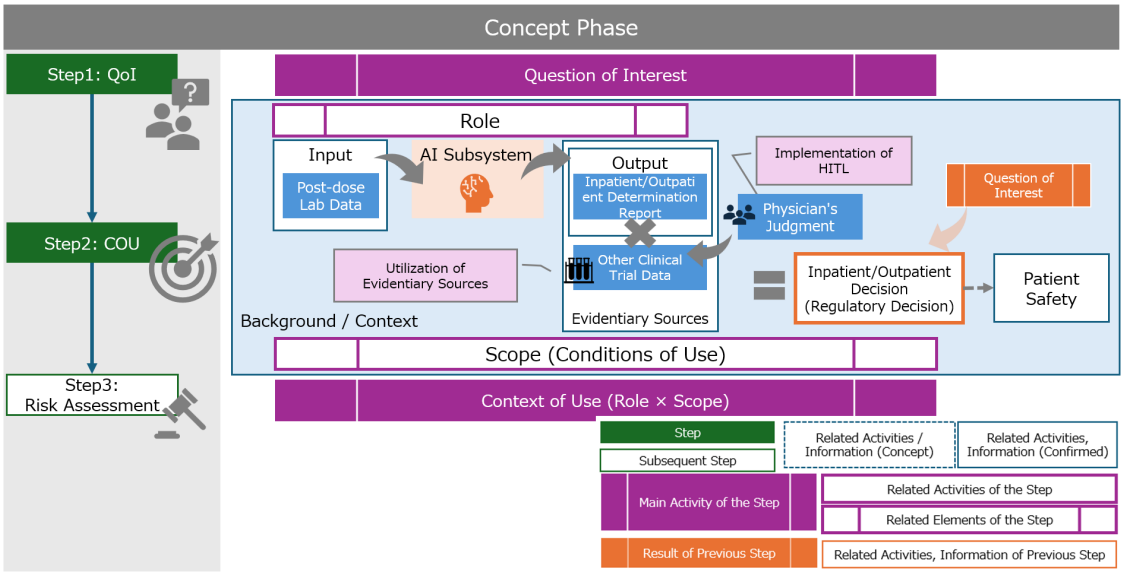

Figure 9: Define the Context of Use for the AI Model

1) Importance of Defining COU

The concept of "Intended Use" in traditional computerized system validation is analogous to the COU of an AI system. However, definitions of "Intended Use" are often limited to the role and scope of the computerized system itself, covering aspects such as the usage environment, required functions, and applicable scenarios. In contrast, GxP processes that can be completed solely on the basis of AI system outputs are extremely rare. In most cases, additional measures are indispensable, such as human review conducted under clearly defined purposes and conditions of use. Defining the COU in Step 2 therefore plays a crucial role in appropriately calibrating the

level of reliance on the AI system within GxP decision-making. It also establishes the COU as a foundation for objective risk assessment and for maintaining accountability with respect to AI system performance and usage processes.

2) Definition of Role and Scope

The COU defines the usage conditions (scope) of the AI system by clarifying not only the intended action using the output (role), but also the relationship between the AI system's output and other evidentiary sources. For instance, in a GCP context where an AI classifies a subject as "low risk," the system's role varies significantly depending on whether the decision to hospitalize relies solely on that output or incorporates a physician's diagnosis. Similarly, in a GMP context, even if the AI system detects filling defects, its role remains relatively limited if an independent sampling inspection is conducted for the final release decision. Thus, the "scope" within the COU, the usage conditions of the AI output, and its combination with other evidentiary sources directly influence how much weight is placed on the AI model's 'role' and how it is positioned within the overall decision-making process."

3) Identification of Evidentiary Sources

Sponsors identify the Evidentiary Sources listed in Step 1, specify them in Step 2, and document how the AI system's output and these Evidentiary Sources are integrated in the decision-making process.

4) Positioning of HITL

HITL is a process where experts review and verify the AI system’s output and are involved in the final decision-making. When defining the scope of the COU, the presence or absence of HITL and the extent of human involvement are particularly important. Annex 22 recommends a process incorporating HITL where human experts finally verify the AI system’s Output. This can reduce the risk that an incorrect Output adversely affects patient safety, product quality, and the reliability of results. However, HITL does not demonstrate a risk mitigation effect in all cases. HITL loses its effectiveness when it becomes a perfunctory confirmation task or when the reviewing expert lacks sufficient information to make a judgment. Therefore, when defining HITL in the COU, practitioners are encouraged to specifically describe the information (Evidentiary Sources) and criteria on which the expert bases the final decision.

5) Documentation of COU (Recommended)

While documenting the COU at this stage is not mandatory, doing so prepares the necessary information for documentation in Step 4. It is recommended that the COU include the following perspectives:

- **Role of the AI model**
  - Input to the AI model (data type, format, acquisition method)
  - Output from the AI model (predicted value, classification result, confidence score, etc.)
  - Contribution of the Output from the AI model to the QoI and decision-making
- **Scope of the AI model**
  - Business process utilizing AI
  - Usage environment (user, location, timing)
  - Usage method of the Output from the AI model (automatic judgment, reference information in HITL, etc.)
  - Other Evidentiary Sources and their utilization/judgment processes
  - Constraints and prerequisites for use

Furthermore, since the definition of the COU serves as the foundation for the entire AI project, ensuring its quality is extremely important. A well-defined COU has the following characteristics:

- **Specificity:** The components of the QoI are clear.
- **Measurability:** It is defined in a way that allows for the performance evaluation of the AI model.
- **Clinical/Operational Validity:** It conforms to the needs of the actual medical or manufacturing site.
- **Feasibility:** It can be addressed with realistic data and AI model technology.

5.3 Step 3: Assess the AI Model Risk

[Overview]

The objective of Step 3 is to assess the AI model risk based on the COU defined in Step 2. This risk is assessed by combining the following two factors:

- **Decision Consequence ("DC"):** The significance of the result if the AI output is incorrect.
- **Model Influence ("MI"):** The degree of influence of the Output from the AI on regulatory decision-making.

AI model risk is determined by the combination of Model Influence (MI) and Decision Consequence (DC), reflecting the degree to which an incorrect AI model output may lead to an Adverse Outcome affecting patient safety, product quality, or the reliability of results.

Table 5 shows specific examples of AI model risk assessment in the GCP and GMP fields based on the cases in Section 5.1 .

Table 5: AI Model Risk Assessment

Case	Decision Consequence	Model Influence	AI Model Risk
GCP: (Investigational Drug A)	**High:** Patient Safety. If the Output is incorrect (i.e., a high-risk patient is misclassified as low-risk), the patient may face a life-threatening situation; thus, the consequence is extremely critical.	**High:** Since the decision for inpatient vs. outpatient monitoring relies solely on the AI Output, the model's influence is maximal.	**High x High = High:** Requires the most rigorous credibility assessment and operation.
GMP: (Injectable Drug B)	**High:** Medication Error. If the drug volume deviates, the correct dose cannot be administered to the patient, potentially leading to medication errors or health damage; thus, the consequence is critical.	**Low:** Verification by separate process exists. Although AI performs 100% inspection, it is positioned as a primary screening. The final release decision involves an independent verification process (risk control measure) via "sampling inspection by personnel." Since AI is not the sole evidentiary source, its influence is limited.	**High x Low = Medium:** Credibility assessment and operation levels can be set rationally.

[Discussion]

Step 3 assesses the potential risk of the AI model. "AI model risk" does not refer to defects, failures, bugs, or security risks inherent to the AI model algorithm itself. It is an assessment of "to what extent subsequent processes will negatively affect patient safety, product quality, and the reliability of results as a result of acting according to incorrect Output from the AI system."

In the example of the GCP field, since the AI system's Output is the sole basis for determining the necessity of hospitalization, the influence is assessed as "High." On the other hand, in the example of the GMP field, since the AI system is a primary screening and there is an independent sampling inspection, the influence is assessed as "Low."

The overall picture of the Concept Phase, associating the activities from Step 1 to Step 3, is shown in Figure 10. In the FDA AI Guidance, AI model risk is assessed along two dimensions: DC, which indicates how serious the consequences would be if the system's Output were incorrect in regulatory activities or decision-making, and MI, which represents the degree of influence of the AI system's Output in regulatory activities or decision-making. The COU from Step 2 serves as the basis for these risk assessments.

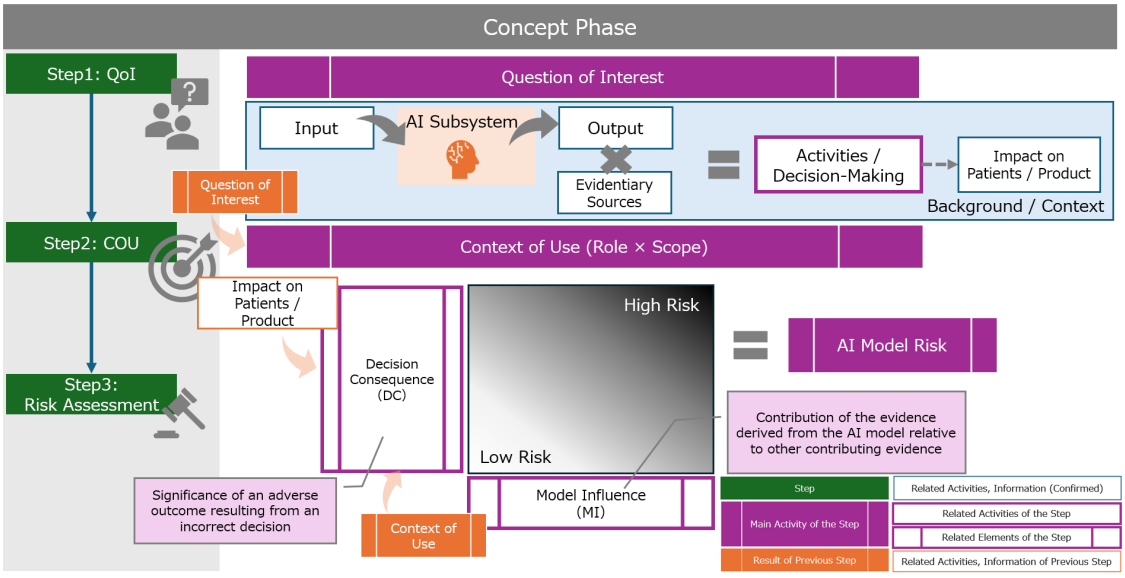

Figure 10: Step 3 Assess the AI Model Risk

Depending on the level of the AI Model Risk, the rigor and scope of the Credibility Assessment (model testing) in the next Step and subsequent operation (monitoring, etc.) are determined. The level of activities commensurate with the risk should be defined in each company's SOPs and equivalent documents.

1) Assessment of Decision Consequence

Errors such as false positives, false negatives, excess, or deficiency can occur in the AI system's Output. When assessing the criticality of DC, rather than simply assessing it as high or low, documenting the evaluation axes and their results based on the COU as the rationale for DC makes it possible to demonstrate the validity of the DC and facilitates impact assessment when the COU is changed.

2) Assessment of Model Influence

In the MI assessment, the "Scope of the AI model" identified in Step 2 becomes a critical factor for judgment. Assessment is performed from the following perspectives:

- Whether the AI Output is the sole basis for decision-making or one of the other Evidentiary Sources.
- Presence or absence of HITL.
- Presence or absence of other independent concurrent processes.

3) Risk Assessment of Dynamic / Adaptive AI Models

The risk assessment in the FDA AI Guidance does not consider the presence or absence of self-learning in AI models. However, as stated in Section 5.7 , AI models that self-learn and adapt to the environment are not excluded. On the other hand, Annex 22 restricts the GxP use of self-learning AI models. Therefore, when using Dynamic / Adaptive AI models that self-learn, it is considered that additional risk assessment items such as Adaptiveness are necessary.

However, since Dynamic / Adaptive AI models are currently restricted under Annex 22, the risk should be assessed as considerably high, and the implementation hurdles are significant. Furthermore, there is a possibility that considerably rigorous Credibility Assessment and related activities will be required even in inspections.

5.4 Step 4: Develop a Plan to Establish AI Model Credibility Within the Context of Use

[Overview]

Step 4 is the planning for the Credibility Assessment of the AI model's Output. A Credibility Assessment Plan tailored to the specific intended use of the AI model is formulated. 4.a and 4.b outline general considerations and evaluation methods regarding the construction and Credibility Assessment activities of the AI model.

4.a: Describe the model and the model development process

i. Describe the model

ii. Describe the data used to develop the model

iii. Describe model training

4.b: Describe the model evaluation process

AI-related technology is evolving rapidly, and the use of AI models in the drug product life cycle is expected to expand further in the future. Therefore, activities to establish the credibility of AI model outputs must generally be commensurate with the COU and risk. Consequently, the content of the Credibility Assessment Plan described in this document should be adjusted accordingly.

[Discussion]

The Credibility Assessment Plan serves as the test plan for the AI model developed through the exploratory development process. A simplified diagram focusing on Step 4 Credibility Assessment Plan from Figure 5 is shown in Figure 11.

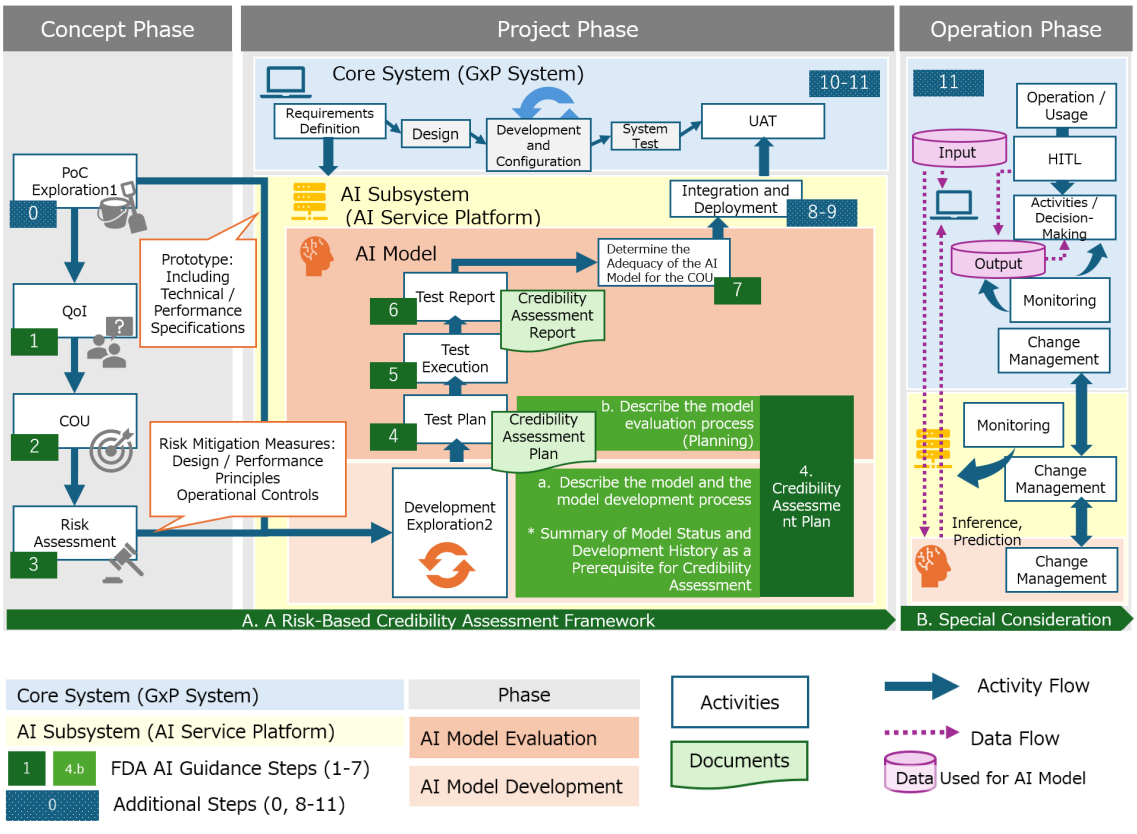

Figure 11: Step 4 Credibility Assessment Plan

Exploratory development is continued based on the technical information from the exploratory development in the PoC of Step 0 and the risk mitigation measures from the risk assessment results

of Step 3. This "exploratory development" corresponds to "Development: Exploration 2" in Figure 11. When the development of the AI model is completed, parameter candidates that satisfy the AI model performance are determined. In this framework, information during development, such as these parameters, is summarized as "4. a Describe the model and the model development process." Based on this information, the "b. Model Testing" process, which is the Credibility Assessment, is planned. In other words, the "Credibility Assessment Plan" is a document that outlines the status and history of the model to be tested and plans the test process to verify its validity.

This concept differs significantly from the concepts of traditional software development and CSV. It is inferred that the FDA intentionally uses the new concept of "Credibility Assessment" instead of using terms like testing or validation. Consistent with this, the FDA AI Guidance does not prescribe a so-called "Validation Plan" for AI model implementation, nor does it address the development process itself. This further reflects how AI model credibility differs fundamentally from conventional CSV.

As the methods and technologies for AI model Credibility Assessment are evolving rapidly, uniform approaches may not be sufficient to address all emerging challenges. The content described in this document is also based on the draft FDA AI Guidance, and practitioners will need to plan with the latest technologies and regulations in mind. For this reason, the FDA strongly recommends discussing and agreeing with regulatory authorities on the scope of AI use and evaluation methods from the planning stage.

The FDA recommends dialogue at the early development stage regarding innovative AI approaches. Especially in cases involving the use of unprecedented new AI technologies, high-risk COUs, or complex HITL processes, early dialogue allows for understanding regulatory expectations and reaching prior agreement on validity using the Credibility Assessment Plan. As a result, the items required by the FDA become the content of the Credibility Assessment Plan, and the risk of regulatory comments during review can be significantly reduced.

The main items required for this plan are summarized in Figure 12.

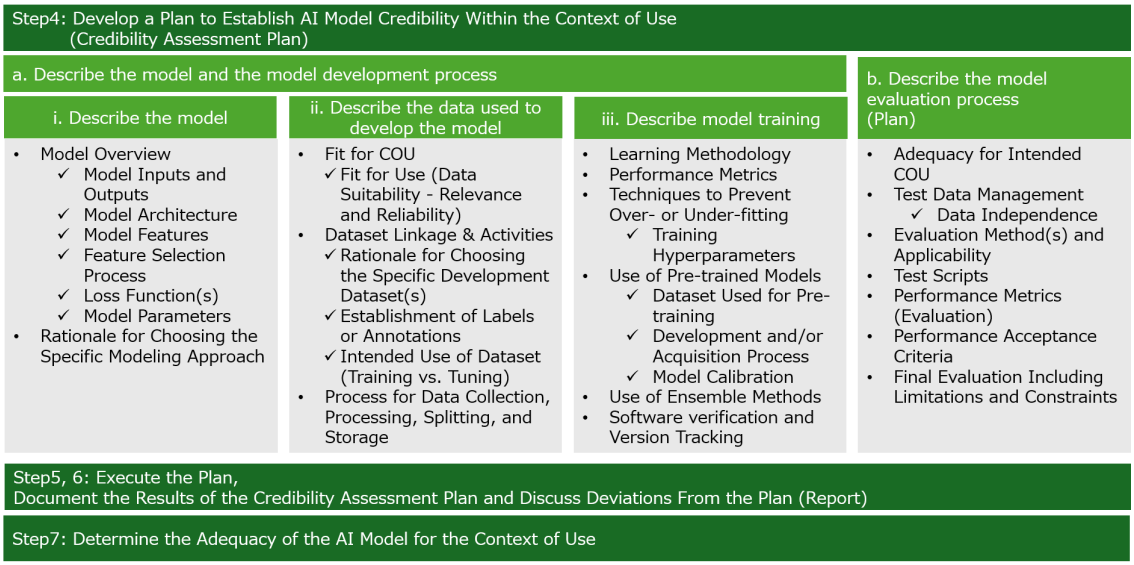

Figure 12: Overview of Step 4 Credibility Assessment Plan

Note: The notation "4.a.i" etc. in this document corresponds to the subsections (4.a.i, 4.a.ii, etc.) of Section 4 in the FDA AI Guidance. It differs from the section numbers in this document.

As indicated in the Computer Software Assurance ("CSA") guidance, the FDA requires that the level and rigor of assurance activities be commensurate with the risk and Intended Use of the computerized system. At the same time, the FDA requires accountability and transparency regarding the rationale for selecting those methods. Practical interpretations of the CSA approach is detailed in a publication from the Japan Pharmaceutical Manufacturers Association (JPMA) "Overview of the FDA CSA Draft Guidance and Examination - Consideration of an Application to the GxP Area" ("Draft CSA Considerations"). In the document, the CSA concepts are extended beyond their original scope in the medical device field, and applied to assurance activities for computerized systems within the GxP domain. Furthermore, a systematic framework for determining the appropriate level and rigor of these activities is demonstrated (see Figure 13.) Although the finalized CSA guidance (Second Edition) has been issued, the draft CSA concepts illustrated in Figure 13 remain unchanged.

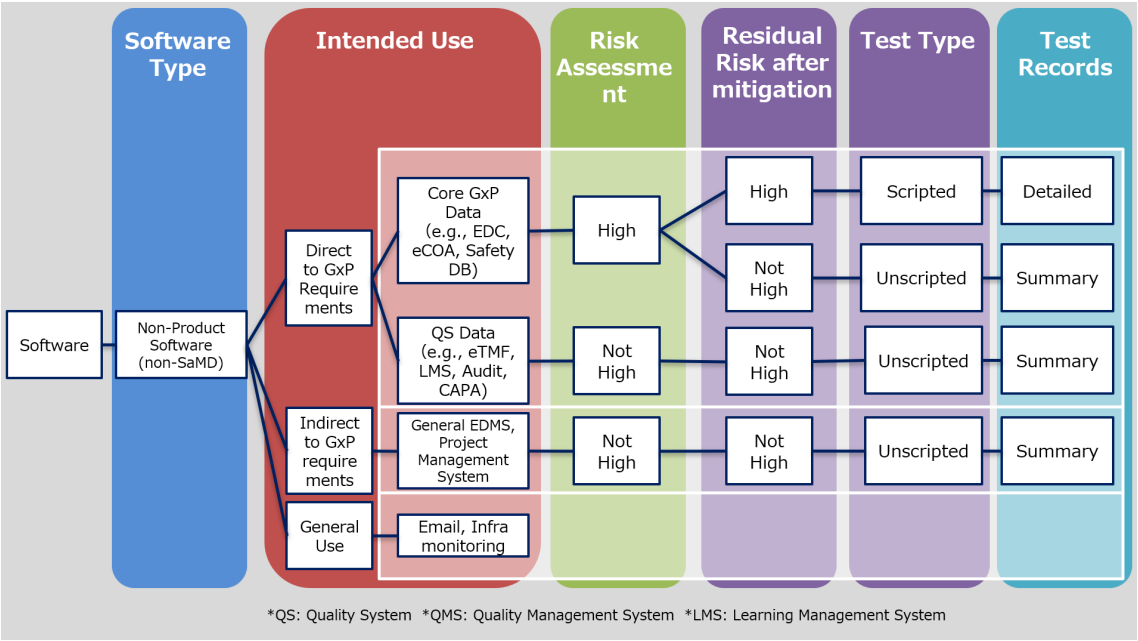

Figure 13: CSA Flow for GCP (Excerpt from Figure 13 of "Draft CSA Considerations", Translated)

This principle applies equally to the Credibility Assessment of AI models. In the "Step 4: AI Model Credibility Assessment Plan," sponsors are required to explain the rationale for each activity and selected method, and furthermore, to explicitly state the limitations and boundaries of the AI model. These requirements are listed as items to be documented, primarily focusing on the performance evaluation of the AI model against the COU.

Regarding Annex 22, when using an AI system in the medicinal product lifecycle, its inference or prediction is required to demonstrate reliability equivalent to or better than existing processes or expert judgment. If the AI system's Output directly affects patient safety, product quality, or data integrity, extremely high reliability is required. Therefore, it is also necessary to evaluate superiority over current processes.

5.4.1 Step 4 a i. Describe the Model

[Overview]

The overview of the developed model should be included in the Credibility Assessment Plan.

- Model Overview
  - Model Inputs and Outputs
  - Model Architecture (e.g., CNN / Convolutional Neural Network)
  - Model Features

- Feature Selection Process and any loss function(s) used for model design and optimization (if applicable)
- Model Parameters

- Rationale for Choosing the Specific Modeling Approach: A concise justification for why the chosen modeling method was selected in light of the problem definition, data structure, and risk level.

[Discussion]

Since an AI system's Output may impact patient safety and product quality, the AI model cannot be treated as a "black box." Therefore, sponsors are responsible for explaining the "logic, development background, and concept of the AI model" in a manner understandable to third parties. The "Model Overview" describes the results of exploratory development, as shown in Figure 14.

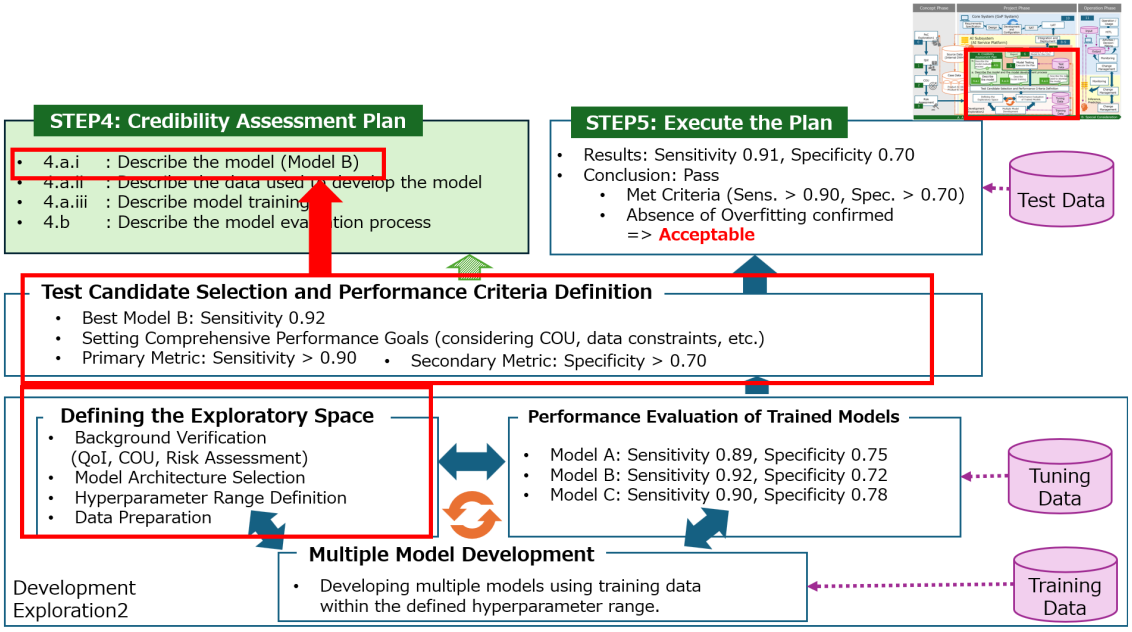

Figure 14: Step 4 a i. AI Model Overview

Note: This figure details the activities ranging from "Exploratory Development (Development Exploration 2)" to Step 5, as presented in Figure 5.

The level of detail in the AI model overview depends on the risk associated with the AI model. For high-risk models, detailed information on all the above items, as well as additional information, may be required. Conversely, for low-risk models, minimal information such as inputs, outputs, and an overview of the architecture may be sufficient.

The FDA specifies the following aspects to be documented.

- **AI Model Overview**
  - Model Inputs and Outputs:
    - Information defined in the Concept Phase (QoI, COU), including the content, media, and pathways of input and output data.
    - Annex 22 requires that AI models performing prediction or classification provide not only the Output but also auxiliary information, such as a "Confidence Score" indicating the certainty of the prediction, and applicable thresholds. This information indicates the extent to which humans can rely on the AI model's prediction for decision-making, thereby ensuring a safer final judgment process.
    - If the Confidence Score is extremely low, generating an Output of "Unpredictable" should also be considered.
  - Model Architecture:
    - The basic structure and type of the AI model (e.g., CNN).
  - Model Features:
    - Specific perspectives, image regions, or variables within the data that the AI focuses on to generate the Output.
  - Feature Selection Process and any loss function(s) used for model design and optimization (if applicable):
    - Feature Selection Process: The rationale and process for selecting final features from numerous candidates (e.g., selected the top 20 features with the highest correlation to the target disease from existing Clinical Trial A and Clinical Trial B).
    - Loss Function: A function to quantitatively evaluate the divergence of the AI model's prediction from the correct answer (ground truth). The goal of training is to define a combination of internal parameters that minimizes the value calculated by this function (loss value). A smaller loss value indicates that the model's prediction is closer to the correct answer (e.g., minimizing false negatives or false positives).
  - Model Parameters:
    - Internal numerical values of the model (such as weights and biases) obtained by learning from training data. Examples include weight matrices in each layer of a neural network, features and thresholds used for splitting each node into a decision tree, and leaf weights in XGBoost. These include internal rules and values determined by learning. These parameters are automatically optimized during the development process and are not directly set by the developer.

- **Rationale for Choosing the Specific Modeling Approach: The basis for selecting the above model and features.**
  - For example, the following perspectives should be included:
    - ✧ Why the selected method is appropriate for addressing the defined QoI and COU
      e.g., rationale for choosing a non-linear over a linear model; rationale for adopting black-box deep learning
    - ✧ Why the selected method is considered appropriate for the data structure
      e.g., a time-series model is suitable because the data is time-series; a tree-based model was selected because the data is structured tabular data
    - ✧ Why the selected method is considered appropriate for this risk level and COU
      e.g., rationale based on explainability, reproducibility, implementation maturity, and existing track record

5.4.2 Step 4 a ii. Describe the data used to develop the model

[Overview]

During the development of an AI model, data is generally split into three types: training, tuning, and testing, and used only for their respective purposes. Training data is used for the development of the AI model (including the definition of model weights, connections, and components). Tuning data is used when performing small-scale evaluations on the trained AI model. Tuning data is used before the final testing phase of the AI model and is part of the AI model development process. Test data is not used during the AI model development.

The performance of an AI model relies heavily on the development datasets (training and tuning datasets) used. Therefore, the development datasets used for the AI model should be "fit for use." This means the data should be both "relevant" and "reliable."

- **Relevant:** Includes key data elements and sufficient numbers of representative participants, or is sufficient data that is representative of the manufacturing process or operation.
- **Reliable:** Is accurate, complete, and traceable.

Commensurate with the model risk, pharmaceutical companies should define data management procedures regarding development datasets and characterize the development datasets. These procedures help identify potential limitations of the data and identify appropriate Credibility Assessment activities to demonstrate the suitability of the AI model for a specific COU.

The summary of the data used for model development is documented in the Credibility Assessment Plan.

- Description of development datasets
  - ➢ Description of development datasets (including how the development datasets were split into training, tuning, and any additional subsets)
  - ➢ Specification of which model development activities were performed using each dataset
- Procedures for collection, processing, annotation, storage, and control of development data, and intended use (training, tuning)
  - ➢ Rationale for choosing the specific development dataset(s)
  - ➢ Procedures for establishing labels or annotations
- Suitability of development data for the COU
- Relevance and reliability of development data
- Control status of development data

[Discussion]

Data management for developing AI models is a unique activity in AI system development. Throughout the entire data management process of "what quality of data," "through what management process," and "how it was used," ensuring transparency based on scientific rationale and maintaining traceability are required. The "Data Overview" describes the data used in exploratory development, as shown in Figure 15.

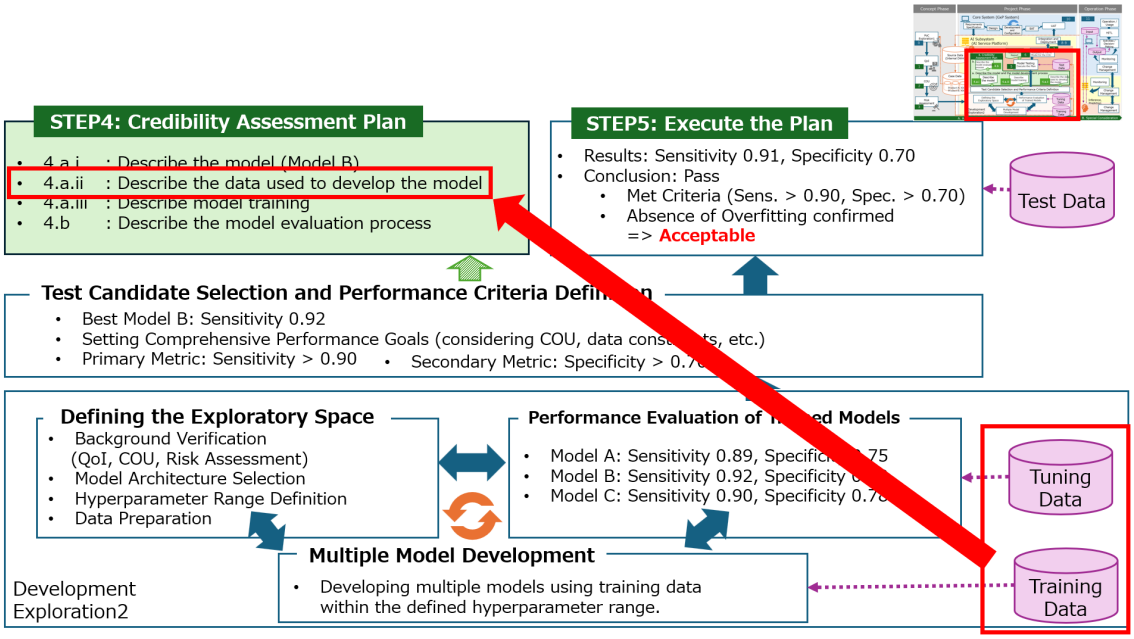

Figure 15: Step 4 a ii. Data Overview

Note: This figure details the activities ranging from "Exploratory Development (Development Exploration 2)" to Step 5, as presented in Figure 5.

Training and tuning data used during development must be strictly separated from test data, and their management must include strict control of access rights. This strict separation demonstrates transparency to ensure the reliability of the model evaluation.

The types and purposes of each data set are summarized in Table 6.

Table 6: Types and Purposes of Data Used for AI Models

Category		Purpose
Development Data	Training Data	Data used to develop (train) patterns in the AI model. This data forms the fundamental performance (knowledge and capabilities) of the model.
	Tuning Data	Data used to evaluate the performance of trained models and to select the specific AI model to be tested.
Test Data		Data used for the final and objective evaluation of the "true performance (generalization capability)" of the model after training and tuning are complete. A fundamental requirement is that this data must be completely independent and unused during the training or tuning processes.

Note: Data referred to as "tuning data" in FDA AI Guidance is called "validation dataset" by Annex 22 and in general machine learning literature . However, in this document, the term "tuning data" is used consistently following the FDA notation to avoid confusion with "validation" in the GxP context (i.e., computerized system validation activities).

The data flow for AI model development and Credibility Assessment (Testing) is shown in Figure 16.

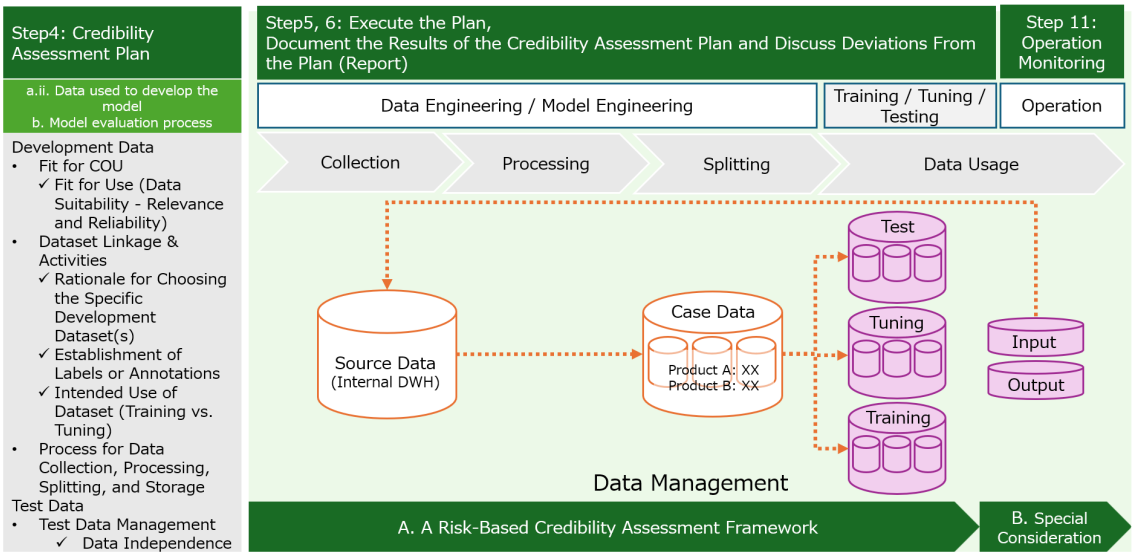

Figure 16: Overview of Data Management for AI Models

Note: The acquisition and management pathways for inputs and outputs vary depending on the configuration of the overall AI system, including the core system. Examples are provided below:

- Databases for inputs and outputs are integrated into the core system.
- Inputs and outputs are maintained in independent databases separate from the core system.
- Inputs are obtained from an internal Data Warehouse (DWH) where various data are collected and integrated. Outputs are also stored in the internal DWH.

Data used for the development and testing of AI models is required to be strictly managed throughout the entire process. The assumed main data management processes are summarized in Table 7.

Table 7: Data Management Processes

Processes	Objective & Activities	Key Requirements
Data Acquisition	✓ **Objective:** To collect and select source data consistent with the AI model's COU. ✓ **Activities:** Identification of data sources, formulation of collection plans, definition of data specifications, and initial quality checks of collected data.	✓ **Source and Rationale:** Clearly state which clinical trials, systems, or batches the data were collected from and the rationale for source selection. ✓ **Relevance to COU:** Explain that the collected data appropriately represents the actual usage environment (e.g.,

Processes	Objective & Activities	Key Requirements
		distribution of race, gender, severity) without bias. ✓ **Consideration of Data Drift:** Explain how risks of divergence between historical data and real-world data (environmental changes/gaps) were considered.
Data Preprocessing	✓ **Objective:** To process source data into a format usable by the model. ✓ **Activities:** Imputation, noise reduction, normalization, annotation (labeling by pattern). Storing and managing case data as datasets per pattern based on annotation information.	✓ **Documentation of Procedures:** Record data processing steps and ensure reproducibility. ✓ **Reference Method:** Define the standard for annotation (e.g., consensus of specialists) and explain the basis for its reliability. ✓ **Annotation Quality:** Document annotator qualifications, SOPs, and Quality Control (QC) processes.
Case Data (Dataset) Splitting	✓ **Objective:** To split data according to purpose. ✓ **Activities:** Formulation and execution of a strategy to split all data into training, tuning, and testing sets.	✓ **Ensuring Independence:** Explain specific methods (e.g., temporal, geographical, or physical separation) to ensure complete independence between development and test data. ✓ **Rationale for Splitting:** Explain the scientific basis for the chosen split ratio and method. ✓ **Justification for Overlap:** If overlap between development and test data occurs, provide a detailed explanation and evaluate the impact on testing.

Processes	Objective & Activities	Key Requirements
Data Governance & Traceability	✓ **Objective:** To maintain and manage data integrity, security, and reproducibility throughout the data life cycle. ✓ **Activities:** Data storage, version control, access control, and audit trail management.	✓ **Storage & Control Procedures:** Methods for data storage location, security measures, and access control management. ✓ **Version Control:** Record and manage accurate version information for datasets, software, libraries, and tools used. ✓ **Quality Assurance:** QA/QC procedures to verify code and calculation accuracy. Ensuring Data Integrity (ALCOA++ principles).

The performance of an AI model relies heavily on the datasets used from model training to testing. Therefore, collecting data that is "fit for use" is required. The "relevance" of data is described based on the examples in Section 5.2.

"Relevance" in the GCP Field Example (Investigational Drug A):

When creating an AI model to assist in the diagnosis of Japanese patients, if only data from populations of European descent is available, the data relevance is determined to be low. Disease trends and physical characteristics may differ depending on race and culture.

"Relevance" in the GMP Field Example (Injectable Drug B):

When creating an AI model for anomaly detection on a manufacturing line, relevance can be considered high if the data covers not only non-defective products but also defective products representing a variety of anticipated defect types (such as chips, stains, and printing errors) in sufficient quantities.

Furthermore, data with sufficient "relevance" and "reliability" must simultaneously statistically represent the data to be processed in the environment where the model is deployed. The possibility of "Data Drift," where past and current data trends diverge, must always be considered, and it must be explained that the dataset faithfully reflects the current COU.

Next, pre-processing and annotation are performed to classify each data so that it can be used for development and testing. The pre-processing process (transformation, normalization, standardization, annotation, etc.) is planned in advance and performed within the scope of satisfying the intended use. Therefore, excessive cleaning or exclusion of data is not recommended, and justifying the validity of each activity is required.

Annex 22 advises against using Generative AI to generate or augment training data. If data using Generative AI is used for development or testing, the legitimacy and validity of such data must also be guaranteed.

Annotation is the task of creating "ground truth data," and its quality significantly affects the AI model's performance. For example, in the case of an AI model that determines "suspicion of cancer" from X-ray photographs, who applied the correct label (annotation) for the "cancer image" becomes an important element. A process where "labels were applied through consultation by two or more experienced specialists" is considered to have significantly higher reliability than a process where "labels were applied by a single resident."

After pre-processing is completed, the collected case data is split according to each purpose (training, tuning, testing). Data splitting is an essential procedure to prevent "Over-fitting," where the AI model excessively memorizes only the training data, and to evaluate generalizable capabilities.

Examples of splitting methods include the following:

- **Temporal Split:** Partitioning data based on a specific point in time (e.g., develop with data up to 2020, test with data from 2021).
- **Geographical Split:** Partitioning data based on distinct sites or regions (e.g., training on data from Site A and testing on data from Site B).
- **Physical Split:** Segregating data based on physical units such as batches or lots.
- **Characteristic Split:** Allocating data to development and testing sets based on specific case characteristics (e.g., product type, disease severity) to ensure balanced representation.

Development data and test data independence is a key principle for ensuring the reliability of model evaluation. However, in the following cases, partial overlap of data or special partitioning methods may be selected as scientifically valid methodologies:

- **Cross-validation:** A method of splitting data into multiple folds and using each fold in turn as a test set.
- **Stratified sampling:** A method of controlling so that all subgroups, including rare events, are appropriately distributed in training and testing.
- **Leave-one-out validation:** A method used when the number of cases is extremely limited.

When adopting these methods, the following should be documented:

- Scientific rationale for selecting the method.
- Data characteristics (number of cases, rarity, etc.).
- Details of the method (partitioning method, number of folds, etc.).
- Impact on performance evaluation (risk of overfitting or overestimation).
  - Overfitting: Risk of overfitting due to using data from the same subject for both training and testing.
  - Overestimation: Risk of overestimating performance by treating verification results used for model selection (tuning) as final performance.
- Risk management measures (conservative threshold setting, additional verification, etc.).

In this section, "Data Management" has been described focusing on development data. The following Sections 5.4.3 and 5.4.4 describe how these data are utilized, explaining the activities of training, tuning, and testing.

5.4.3 Step 4 a iii. Describe the Model training

[Overview]

The summary of the model development results is included in the Credibility Assessment Plan.

- Model Training Methodology
  - Learning methodology (e.g., supervised, unsupervised).
  - Performance metrics used to evaluate the model. All performance estimates are provided with confidence intervals. Examples of performance metrics are shown below.
    - ✧ Receiver Operating Characteristic (ROC)
    - ✧ Recall or Sensitivity, Specificity, Positive/Negative Predictive Values (PPV/NPV), True/False Positive and True/False Negative counts (e.g., in a confusion matrix)
    - ✧ Positive/Negative Diagnostic Likelihood Ratios (PLR/NLR), Precision, F1 scores
  - Techniques employed to prevent over- or under-fitting (e.g., regularization techniques).
  - Training hyperparameters (e.g., learning rate, regularization coefficient, batch size). Unlike hyperparameters, the loss function is a structural element selected during the model architecture design phase (e.g., cross-entropy loss, Mean Squared Error) and is treated as distinct in this document.
  - Presence or absence of the use of a pre-trained model (or multiple pre-trained models).
- Use of Pre-trained Models (or multiple pre-trained models)
  - Datasets used for pre-training and the development and acquisition methods of the pre-trained model.
- Use of ensemble methods.

- AI Model calibration
  - ➢ Fine adjustment aimed at improving the accuracy and/or repeatability of the Output of the trained model.
- Quality Assurance and Control Procedures for Computer Software (including tools and packages constituting the development environment) and version history.

[Discussion]

This section of the FDA AI Guidance requires documentation of the model submitted for evaluation testing, but does not specify how that model should be developed. This ambiguity is what makes Step 4 a iii. one of the more challenging sections for practitioners to interpret. To aid understanding, this document uses grid search as a working example of exploratory and iterative model development (see Figure 17 and Figure 18). The same principles apply to other hyperparameter optimization approaches such as random search and Bayesian optimization.

Exploratory development of AI models involves creating and comparing multiple model candidates in parallel to identify the optimal solution. The activities and outcomes of this development process should be documented as a "Summary of Training and Tuning" in the "Credibility Assessment Plan."

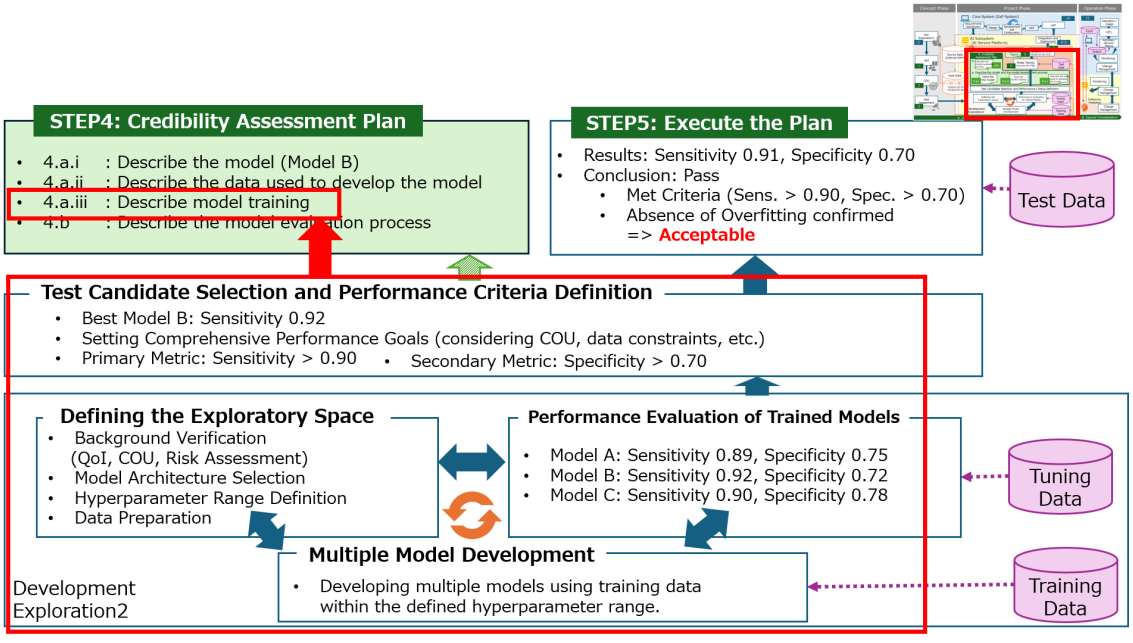

Figure 17: Step 4 a iii. Describe model training

Note: This figure details the activities ranging from "Exploratory Development (Development Exploration 2)" to Step 5, as presented in Figure 5.

As illustrated in Figure 17, exploratory AI model development involves generating multiple candidate models in parallel, comparing their performance, and selecting the optimal solution. This process and its outcomes are then compiled into the "Credibility Assessment Plan."

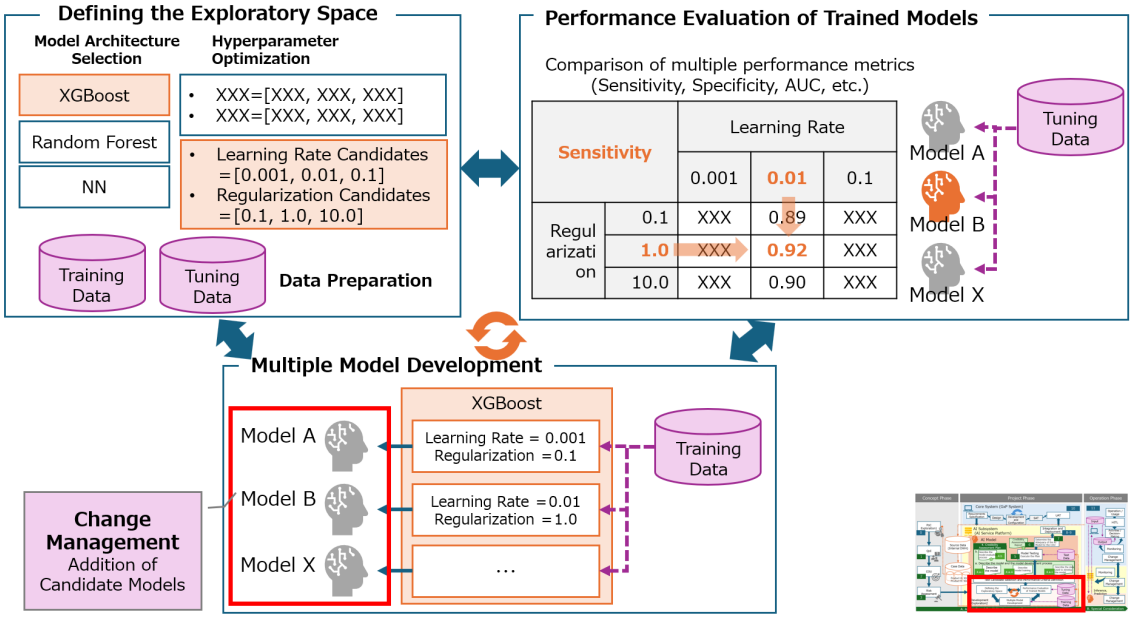

Figure 18: Exploratory and Iterative Training and Tuning Process

Note: This figure details the activities of "Exploratory Development (Development/Exploration 2)" as presented in Figure 5.

Figure 18 illustrates an example of an exploratory development process based on grid search hyperparameter optimization, in which predefined ranges and step sizes for parameters such as learning rate and regularization coefficient are systematically evaluated across all combinations to train and assess multiple models. A key characteristic of AI model development is that activities during this stage, including training and tuning, are treated not as "changes" in the traditional change management sense, but as the "addition of candidate models."

The concepts of change management corresponding to Step 4, as shown in Figure 11, are presented below:

a. Describe the model and the model development process

- Construct and evaluate multiple model patterns in parallel
- Record each pattern as an independent model
- Manage as "addition of model candidates" rather than traditional "changes"
- Finally, select the pattern demonstrating optimal performance

b. Describe the model evaluation process

- Fix and document the specifications of the selected model
- Apply traditional change management processes to subsequent changes
- Retraining and parameter adjustments are subject to change management

As illustrated above, in exploratory development, each iteration is managed as a model record rather than as a version upgrade to a single system.

High-quality training data alone does not guarantee reliable model performance. Flaws in the training method can undermine results regardless of data quality. Selecting appropriate performance evaluation metrics requires statistical and data science expertise and should be done in consultation with relevant specialists. The development history of the AI model, including how it met the target performance criteria, is documented in the Credibility Assessment Plan. This section covers the key principles of AI model training and its quality assurance.

1) Model Learning Methodology

The training process, which fundamentally supports model performance and credibility, is designed based on consistent scientific rationale, from the selection of the approach to performance evaluation and ensuring generalization capability. The model training method is determined by the COU and available data. Major training methods include "Supervised Learning" and "Unsupervised Learning."

- **Supervised Learning:** A method of training the relationship between Input and Output using a dataset with correct labels attached. It is adopted in COUs where clear correct answers exist, such as diagnosis support and defective product detection.
- **Unsupervised Learning:** A method where the model itself discovers structures and patterns (such as clustering) latent in the data from data without correct labels. It is adopted in COUs where clear correct answers do not exist, such as identifying candidate subjects from unstructured data like physician notes in medical records, or discovering similar pathology subgroups through clustering of patient background data.

2) Performance Metrics

Model performance is insufficiently evaluated using only a single metric (e.g., accuracy rate). Especially in issues where the risks of false negatives (e.g., overlooking Disease A) and false positives (e.g., misdiagnosing a healthy person as having Disease A) are asymmetric, multifaceted evaluation is essential. Criteria for these performance evaluation metrics are established at each stage of development and testing.

- **Confusion Matrix:** A table contrasting the model's prediction results with the ground truth, serving as the basis for performance evaluation metrics. This allows for the objective grasp of the numbers of True Positives, True Negatives, False Positives, and False Negatives.

Table 8: Confusion Matrix Table

		Prediction	
		Positive	**Negative**
Actual Condition	Positive	True Positive: TP	False Negative: FN
	Negative	False Positive: FP	True Negative: TN

- **Major Performance Metrics:** Based on the confusion matrix, metrics such as Sensitivity/Recall, Specificity, Precision, and F1 score are calculated. These metrics are often in a trade-off relationship, and practitioners are expected to explicitly state the rationale for which metrics are emphasized according to the intended use of the model.

Table 9: Examples of Performance Evaluation Metrics

Metric	Formula	Purpose
Sensitivity / Recall	$\frac{TP}{TP+FN}$	Indicates the ability to correctly identify positive cases without omission (minimizing false negatives). Critical for safety-related screening.
Specificity	$\frac{TN}{TN+FP}$	Indicates the ability to correctly identify negative cases without misclassification (minimizing false positives).
Precision	$\frac{TP}{TP+FP}$	Indicates the proportion of actual positive cases among those predicted as positive. High precision implies high reliability of a positive prediction.
F1 Score	$\frac{2*(Precision*Recall)}{(Precision+Recall)}$	The harmonic mean of Precision and Recall. Provides a balanced view of performance, especially when datasets are imbalanced. A value closer to 1 indicates better combined performance.

- **Confidence Intervals:** An interval indicating the statistical variability of the calculated performance evaluation metrics. By listing confidence intervals in addition to point estimates (e.g., Sensitivity 95%), the robustness of the comprehensive evaluation results of the Output generated by the AI model can be demonstrated.

1) Model Optimization and Establishment of Performance Metrics

a) Techniques to prevent overfitting / underfitting

AI models need to possess "generalization capability" to demonstrate stable performance not only on training data but also on unknown data. Generalization capability allows for the development of AI that demonstrates quality and performance capable of withstanding real-world environments. As a technique for this, "Regularization," which imposes a penalty on model complexity, is often used. In the Credibility Assessment Plan, practitioners are expected to specifically describe what measures were taken to prevent overfitting.

Furthermore, hyperparameters that control the model training process, such as learning rate and regularization coefficient, significantly influence model performance. It is necessary to systematically explore and adjust (tune) these parameters and document the rationale for selecting the finally adopted values.

b) Determination of Performance Metrics and Acceptance Criteria

Performance metrics and their acceptance criteria are determined based on the COU through the following process:

- Clarification of COU Requirements (Completed in Steps 1-3)
  - Based on the COU defined in Steps 1-3, clarify the "priority" of performance requirements. Examples include "minimizing false negatives is the top priority (emphasis on sensitivity)," "ensuring fair performance across subgroups," or "quantifying prediction uncertainty." This priority serves as the criteria for selecting performance metrics.
- Examination of Candidate Performance Metrics
  - Examine multiple candidate performance metrics that could satisfy the COU requirements. Examples include Sensitivity, Specificity, F1 Score, AUC (Area Under the Curve), and Balanced Accuracy. Clarify how each metric corresponds to the COU requirements.
- Confirmation of Data Characteristics
  - Grasp the characteristics of the actual development data (e.g., prevalence/positive rate, sample size, degree of data imbalance, subgroup distribution). These

characteristics influence the selection of performance metrics for statistically reliable evaluation.

- Exploratory Evaluation and Optimization
  - Trial combinations of multiple model architectures, data pre-processing methods, and performance metrics to identify the combination that best satisfies the COU. During this process, the following should be recorded:
    - ✧ Candidate performance metrics examined (e.g., Sensitivity alone, F1 Score, combination of Sensitivity + Specificity, etc.)
    - ✧ The degree to which each performance metric satisfies COU requirements
    - ✧ Achievability under data constraints
    - ✧ Rationale for selecting the final performance metrics (including combinations)
    - ✧ Finalization and Documentation of Metrics and Thresholds
  - Document the following regarding the finally selected performance metrics and acceptance criteria:
    - ✧ Selected metrics (e.g., Primary metric: Sensitivity ≥ 95%, Secondary metric: Specificity ≥ 70%)
    - ✧ Reason why the adopted performance metrics were evaluated as optimal for the COU (logical consistency with COU)
    - ✧ Reasons for not adopting other candidates
    - ✧ Rationale for setting thresholds (clinical acceptance limits, data characteristics, statistical validity, etc.)
    - ✧ Consideration of confidence intervals
    - ✧ Review results by statistical experts or data scientists

c) Important Principles

This approach accommodates the exploratory nature of machine learning while balancing fidelity to the COU and scientific rigor. The crucial point is not to "arbitrarily select metrics," but to "logically derive optimal metrics under COU requirements and specific data constraints, and to ensure transparency and documentation of that process."

In particular, prior consultation with regulatory authorities is strongly recommended regarding the validity of performance metric selection in the following cases:

- COU with a risk assessment result of "High" (DC: "High" × MI: "High")
- Novel AI methods
- Cases involving complex HITL (Human-in-the-Loop) processes
- Cases where performance differs significantly between subgroups of the dataset

2) Use of Pre-trained Models

Utilizing existing AI models rather than developing models in-house is efficient; however, sponsors are required to clarify the rationale justifying the use of the relevant AI model. The method of using a pre-trained model as a foundation and performing additional training (fine-tuning) with in-house data for a specific task is widely adopted. However, when using pre-trained models in the GxP area, risk assessment including the data used for the relevant pre-trained model and version control is necessary.

When using pre-trained models, clarifying the following information is necessary to maintain the transparency of the AI model; however, consultation with the FDA is required.

- Name and version of the pre-trained model used.
- Source of the model (official registry, etc.).
- Information regarding the characteristics and bias of the dataset used by the original model for training.

3) Ensemble Methods

As one method to improve Output quality, there is the ensemble method, which aims for higher accuracy and robustness than a single model by combining multiple different models and making a comprehensive judgment based on the prediction/inference of each model. If this is adopted, it is clarified what models were combined and by what method (e.g., majority vote, averaging).

4) Final Model Adjustment and Quality Assurance

The model that has completed training requires final adjustments for practical use and quality assurance of the entire development process.

- **Calibration:** A process of fine adjustment to realize the individual "Confidence Score" (e.g., positive with 90% probability) output by the model. By performing appropriate calibration, the reliability probability improves, enabling Output to be more aligned with the actual usage environment such as clinical settings.
- **Software Quality Assurance and Reproducibility:** The reliability of an AI model depends heavily on the reproducibility of its process. Therefore, the environment and tools used for development are recorded and managed.
  - **Recording Development Environment:** Record the configuration and versions of the programming language (e.g., Python), major libraries, and related tools themselves.
  - **Version Control:** To track the change history of source code, configuration files, and execution procedures, a version control system is utilized, and its management

structure is demonstrated. Also, fixing and recording the random seed is essential for ensuring model reproducibility. It is desirable to record the random seed settings used in the framework (e.g., Python, TensorFlow®) in the development records along with other parameters.

5.4.4 Step 4 b. Describe the model evaluation process

[Overview]

This section outlines the testing of the fully trained model. Testing of the AI model is an activity to evaluate whether the model's performance is adequate for the COU using test data. Test data must be independent of development data and must not be used during development. Test data is used to evaluate the AI model's performance after training. Like development data, these data are also required to be "fit for use."

Test-related information to be included in the Credibility Assessment Plan is as follows:

- Procedures for collection, processing, annotation, storage, and control of test data, and the intended use of the AI model.
  - Independence of development data and test data:
    - Data independence can be achieved using data from a different clinical trial or health care system, or data acquired using different batches or products.
  - If any data overlap occurs between development and testing, explain how the data was used and justify why such use was appropriate.
  - As relevant, the reference method used to create the test data, and a summary of the reference method's performance.
- Status of conformity of test data to the COU:
  - If prediction models are developed using historical development data, the AI model may not perform as well in the COU if the development data are different from the data encountered in the deployed environment used in the COU (Data Drift).
- Description of the agreement between the model prediction and the observed data.
- Rationale for the chosen model evaluation method(s):
  - Explain the applicability of the evaluation methods to the modeling method used and to the COU. If the COU involves a HITL, ensure that the evaluation methods consider the performance of the human-AI team, rather than just the model's performance in isolation.
- Clearly describe the rationale for the chosen and planned "test methods and performance metrics." Furthermore, explain the appropriateness of the evaluation methods to the

"modeling method" used during development and to the "COU." Metrics to be considered include:

  - Receiver Operating Characteristic (ROC)
  - Recall or Sensitivity, Specificity, Positive/Negative Predictive Values (PPV/NPV), True/False Positive and True/False Negative counts (e.g., in a confusion matrix)
  - Positive/Negative Diagnostic Likelihood Ratios (PLR/NLR), Precision, and/or F1 scores
  - Process by which the uncertainty and confidence level of model predictions will be estimated.
- Limitations of the modeling approach, including potential biases.
- Quality assurance and control procedures for code verification:
  - Including resolution of any errors or anomalies (e.g., ensuring codes are error-free, calculations are accurate).

[Discussion]

Unlike traditional CSV, which follows pre-fixed specifications, AI model development is exploratory. The model that ultimately meets the target performance criteria becomes the specification. The more exploratory the development process, the greater the objectivity required in the final evaluation. This is why both the FDA AI Guidance and Annex 22 place strong emphasis on the complete independence of test data.

In this light, the final gate of the AI model development life cycle is the objective performance evaluation, i.e., testing, of the completely developed model. This process verifies whether the knowledge acquired by the model through training possesses "generalization capability" to function effectively on unknown data. This is the most scientifically critical process to evaluate whether the model, after training and tuning, can stably demonstrate the intended performance in the actual usage environment. The Credibility Assessment Plan is structured around this activity.

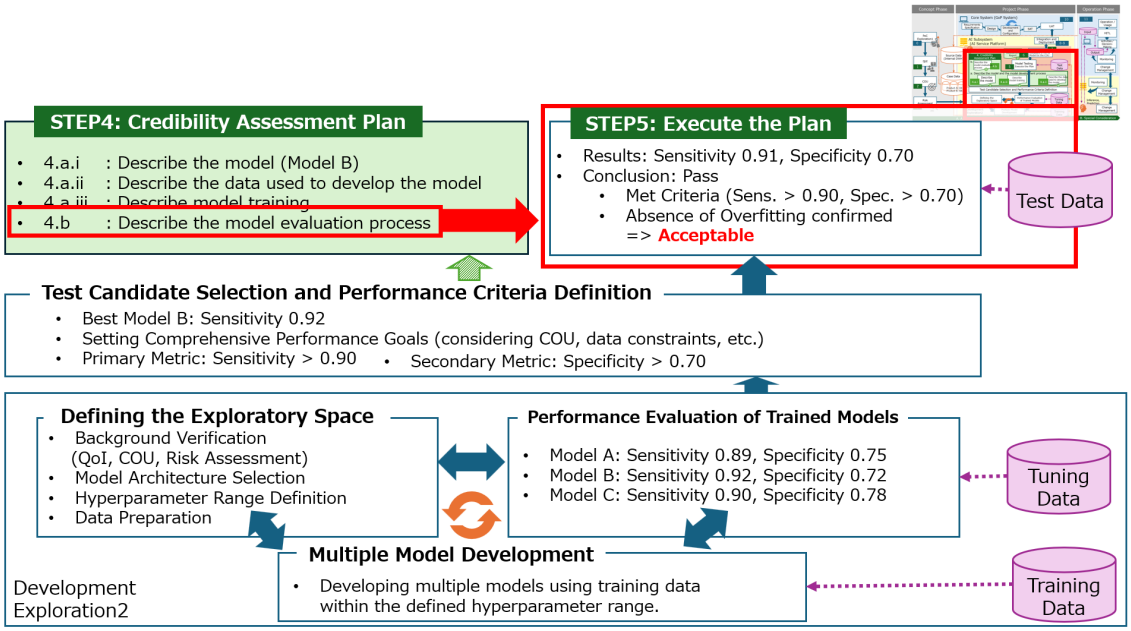

Figure 19: Step 4 b. Describe the model evaluation process

Note: This figure details the activities ranging from "Exploratory Development (Development Exploration 2)" to Step 5, as presented in Figure 5.

Similar to the development of a standard Core System, the Credibility Assessment Plan, which serves as the test plan, is approved before the start of testing for the AI model. It is assumed that the test will include the following perspectives:

- COU, Intended Use
- Management of test data
  - Independence of test data
  - Independence of testing personnel
  - Retention of audit trails including reference to test data
  - Usage method of test data
- Validity of evaluation methods
  - AI utilization process including human involvement and test scope
  - Uncertainty of AI prediction results, estimation process of confidence level
  - Presentation of AI features
- Test scripts
- Performance evaluation metrics
  - Evaluation of generalization capability
  - Validity of performance metrics utilized
  - Calculation methods

- Acceptance criteria
- Final evaluation
  - Limitations and restrictions of AI, etc.

This section details the key points of Credibility Assessment in the final evaluation of the AI model, specifically the following three points:

1) Requirements and management of test data
2) Validity of evaluation methods
3) Performance evaluation metrics

1) Requirements and Management of Test Datasets

As stated in Section 5.4.2, the reliability of the evaluation depends heavily on the quality and management structure of the test data used. The testing is explained focusing on "Independence of Test Data."

- **Ensuring Data Independence:** FDA AI Guidance and Annex 22 both require ensuring the independence of the data itself.
  - Test data must be completely separated from model training and tuning data (prevention of Data Leakage).
  - Independence is ensured through temporal, geographical, or physical splitting.
  - Access control and audit trails in the database.
- **Ensuring Personnel Independence:** Annex 22 also requires the independence of personnel accessing the data. Ideally, development personnel and testing personnel should be completely separated, and this should be demonstrable via access rights to data and audit trails. However, in cases where it is difficult to completely separate personnel due to organizational scale or resources, practical independence and transparency can be ensured by implementing measures such as recording and approving access to test data, conducting third-party reviews (4-eyes principle) during parameter changes or model redevelopment, and retaining chronological records of development and testing activities.
- **Handling of Overlapping Data Use:**
  - If there is an overlap between development and test data, the test fails to measure the true ability of the model and becomes merely a "confirmation of memorization" of the training content, carrying the risk of overlooking Over-fitting. Such overlap includes using test data for training or tuning after evaluating it as "Fail" in the test.

- If the amount of available data is limited, overlapping use of development data and test data may be unavoidable. In such cases, as shown in Section 5.4.2 , documentation such as "scientific rationale for selecting the method" is required.

- **Usage History Management of Test Data:**
  - Annex 22 requires recording which data was used for testing, when (specific information of the AI model such as development pattern information), and how many times, to demonstrate the validity of the test. The usage status of test data includes situations where test data was "not used." If prepared test data was not used, it is treated as a deviation from the test, and the impact on the result is evaluated.

2) Validity of Evaluation Methods

Evaluation methods must be adapted to the characteristics of the model, the actual usage environment, and the COU.

- **Utilization of Explainability**
  - When evaluating an AI model, visualizing features (SHAP values, LIME, heat maps, etc.) and reviewing what basis the AI model used to predict the result serves as evidence to ensure that the model is making decisions based on relevant and appropriate features and based on risk. Where applicable, feature attribution methods are utilized for the explainability of the AI model.
- **Comprehensive Evaluation Including HITL**
  - If the COU assumes "HITL," the evaluation method evaluates not only the model's performance alone but also the human and AI system comprehensively. Annex 22 requires that, at a minimum, when implementing an AI system, performance is equal to or better than the process not utilizing the AI system. In other words, sponsors will need to compare the performance of the process being replaced by the AI model.
  - In the case of a system where a physician makes the final decision, such as a diagnosis support AI system, practitioners will need to evaluate not only the AI model's performance in isolation, but also whether the **overall decision-making process**, including the physician's final judgment, performs at least as well as the process without AI support. Since there is a risk of inducing human error by using the AI system, the comprehensive operational process and usage restrictions of the AI system are considered, including human processes.
  - It is considered acceptable to conduct this evaluation in the UAT of the Core System.
- **Quantification of Confidence Score and Uncertainty**
  - Quantitatively indicating the "Confidence Score" of model prediction is extremely beneficial for users using the Output to interpret it appropriately and prevent over-

reliance. When referencing the AI system's Output that also appends a Confidence Score indicating the degree of confidence in the prediction result (e.g., Positive with 99% Confidence Score, Positive with 50% Confidence Score), the physician's awareness when making a final decision based on the prediction result becomes completely different.

- If there are restrictions such as thresholds for obtaining information necessary for AI prediction, appropriate threshold settings are established so that Output can be output under appropriate conditions. In cases where the Confidence Score is extremely low, outputting "Unpredictable" as an Output rather than making a prediction with low reliability should also be considered.

3) Performance Evaluation Metrics

Model performance is evaluated using multiple metrics considering the risk in the COU. By planning evaluation metrics and acceptance criteria in advance and satisfying these metrics, etc., it can be evaluated that the model performance is suitable for the COU and whether it is acceptable. Where the COU covers distinct subpopulations or product categories, such as inpatient versus outpatient, or Product A versus Product B, acceptance criteria should be defined separately for each case.

Also, for high-risk COUs, in addition to comprehensive performance evaluation metrics, it is recommended to confirm performance metrics and confidence intervals for major patient subgroups or manufacturing conditions, and to evaluate whether performance is significantly inferior in specific populations or conditions (i.e., check for bias or local weaknesses).

The perspectives for major performance evaluation of AI model testing are described below.

- **Evaluation of Generalization Capability:** Evaluate whether the model adapts to the usage environment and possesses "high generalization capability" (i.e., sufficient performance on new data).
  - Detection of over-fitting and under-fitting risks.
  - Ensuring transparency by disclosing known model weaknesses, limitations, and potential biases (e.g., performance degradation in specific patient groups).
- **Performance Metrics:** Present performance evaluation metrics suitable for this COU, such as confusion matrix, Sensitivity, Specificity, Precision, and F1 score (see Section 5.4.3).
  - The specific method for calculating sample size varies depending on the COU, model type, and performance metrics. In practice, it is advisable to consult with a statistician, referencing the frequency of events or event counts obtained from training data (e.g., number of positive cases, number of defective products), to estimate the approximate

number of cases/samples required for test data based on the "expected performance value" and "acceptable margin of error and confidence level".

- **Confidence Intervals:** All performance estimates are accompanied by confidence intervals indicating statistical variability to ensure the reliability of the results (see Section 5.4.3).

In addition, from the perspective of Credibility Assessment, the reliability of any tools or software used for evaluation must be ensured.

- **Quality Assurance of Evaluation Code:** Ensuring the quality of the program code used for evaluation calculations is mandatory. It is necessary to clearly explain the verification procedures, such as code reviews and unit testing, to demonstrate that the code is free of errors and that calculations are accurate. This ensures the reliability of the evaluation results.

Particularly when implementing LLMs or Generative AI to support regulatory decision-making, practitioners will need to verify credibility within the relevant COU through a systematic review by experts of Outputs corresponding to representative prompts and Inputs, and through confirmation of reproducibility by fixing prompts and model versions.

5.5 Step 5: Execute the Plan, Step 6: Document the Results of the Credibility Assessment Plan and Discuss Deviations From the Plan

[Overview]

Step 5 involves the execution of the Credibility Assessment Plan developed in Step 4, and Step 6 involves the documentation of the execution results. In Step 6, the results of the planning and execution from Steps 1 to 4, including any deviations, are compiled into a report. This Credibility Assessment Report serves as documented assurance that an AI model suitable for the COU has been developed. It is important to note that the FDA may request this Credibility Assessment Report.

[Discussion]

Following the development of the "Credibility Assessment Plan (Section 5.4)," the execution of the plan and the documentation of results are performed as the final stages of the AI model project phase (see Figure 11). Steps 5 and 6 comprise activities designed to demonstrate, with objective evidence, that the AI model is credible for its intended purpose (COU). The FDA AI Guidance emphasizes the importance of consulting with regulatory authorities regarding the plan content and reporting methods prior to commencing these activities to ensure alignment on expectations.

1) Execution of Credibility Assessment Plan

Sponsors execute the Credibility Assessment Plan in accordance with "b. Describe the model evaluation process" (model testing including data management, model evaluation, and performance evaluation) as described in the Credibility Assessment Plan. Should any deviations from the plan occur during execution, all details and reasons must be recorded.

- **Consultation with Regulatory Authorities**

  The FDA AI Guidance suggests that engaging with regulatory authorities prior to executing the plan is beneficial for both parties. Through such prior consultation, the following benefits can be obtained:

  - **Alignment of Expectations:** Expectations regarding the validity of the planned evaluation activities (i.e., whether they are commensurate with model risk and the COU) can be aligned with regulatory authorities. This can significantly reduce the risk of rework in later stages and minimize risks pointed out during the marketing application review.
  - **Identification of Potential Challenges:** Potential challenges that the company alone may not be able to fully foresee can be identified, and discussions on how to address them can be held with regulatory authorities at an early stage.

2) Documentation of Execution Results and Discussion of Deviations

The purpose of documentation is to systematically document the results of activities executed according to the plan into a "Credibility Assessment Report." This report is the deliverable of the series of activities concerning the AI model's credibility.

This report is required to primarily include the following content:

- **Summary of Evaluation Results:** Objectively describe all results obtained in Step 5.
- **Discussion of Deviations from the Plan:** If procedures or methods differed from the plan, clearly describe the facts, reasons, and the impact of such changes on the evaluation results. From the perspective of ensuring credibility, it is crucial not to conceal deviations but to logically explain their validity.
- **Reference to Relevant Information:** The report should clearly state the prerequisites for the evaluation, including a summary of the QoI, COU, and model risk assessment results established during the planning phase.

The submission strategy for the Credibility Assessment Report to regulatory authorities differs depending on the AI model risk and the type of application. For example, the report may be (1) a self-contained document included as part of a regulatory submission or in a meeting package, or (2)

held and made available to FDA upon request (e.g., during an inspection). It is recommended to agree on this submission strategy with regulatory authorities prior to the execution of the Credibility Assessment.

5.6 Step 7: Determine the Adequacy of the AI Model for the Context of Use

[Overview]

Based on the results documented in the Credibility Assessment Report, if the AI model is evaluated as appropriate for the specific COU, operations will commence. However, it is possible that the model may be evaluated as not appropriate for the specific COU. Even if the credibility of the model is determined to be insufficient, multiple countermeasures can be considered depending on the risk.

The FDA AI Guidance presents five alternative options:

(1) the sponsor may downgrade the model influence by incorporating additional types of evidence in conjunction with the evidence from the AI model to answer the question of interest;
(2) the sponsor may increase the rigor of the credibility assessment activities or augment the model's Output by adding additional development data;
(3) the sponsor may establish appropriate controls to mitigate risk;
(4) the sponsor may change the modeling approach; or
(5) the sponsor may consider the credibility of the AI model's output inadequate for the COU; therefore, the model's COU would be rejected or revised in an iterative fashion.

[Discussion]

The following examples illustrate how each countermeasure applies when a model's credibility is found to be insufficient.

Table 10: Countermeasures when AI Model Credibility is Insufficient

Countermeasures	Examples
(1) Incorporation of additional Evidentiary Sources	Do not rely solely on AI system predictions; instead, treat AI system output results as reference information.
(2) Increasing the rigor of Credibility Assessment activities or adding development data	To increase the rigor of Credibility Assessment, expand from a single metric to multiple performance metrics to improve the quality of evidence. Alternatively, to improve the model's generalization capability, add data

Countermeasures	Examples
	from similar trials or patient registries and perform re-training.
(3) Establishing appropriate controls to mitigate risk	Continue using the AI model as is, but supplement its performance with additional human processes, such as increasing the frequency of monitoring in the operational process.
(4) Changing the modeling approach	Rebuild the model using fundamentally different technologies or design philosophies.
(5) Rejecting the model's COU or modifying it in an iterative fashion	Discontinue the use of this AI model, or narrow the scope of the COU to limited usage.

The FDA does not assume that AI models will always output perfect results. Therefore, a failure in initial implementation does not signify the end of the project; rather, the guidance preserves a pathway for continuous development. It is crucial to correctly understand the AI model's limitations. If credibility is insufficient, the sponsor should manage risks regarding how to assure patient safety, product quality, and reliability of results, and then reflect these measures in re-development efforts or operational controls.

5.7 Step 8: AI Model Implementation, Step 9: Model Integration and Deployment

[Discussion]

Although the activities in this section are not explicitly detailed in the FDA AI guidance, Step 8 and Step 9 have been added to this document based on the system life cycle concepts of GAMP 5 2nd Edition. These steps involve implementing the AI model within the AI subsystem and establishing connectivity between the AI model and the GxP system (core system). If input data exported from the source system cannot be utilized by the AI model in its raw format, the AI subsystem may include functions to process or transform this data.

Since the primary function of the AI subsystem is data exchange, the main activities are expected to focus on the configuration and verification of network connections and data transmission.

In the construction and operation of the AI Subsystem, the following practical considerations are included:

- Definition of Interface Specifications
  - Input data format, range, handling of missing values
  - Output data format, presence or absence of Confidence Score

- Error handling methods

- Configuration Management of the AI Subsystem
  - Model version management (model file, weights, settings)
  - Fixing libraries and dependencies
  - Ensuring reproducibility of the inference environment:
    - Container technologies such as Docker® are primarily used to stably reproduce the inference environment. By packaging the OS, libraries, and runtime required for inference into a single container image, it becomes possible to reproduce virtually identical inference environments even when deployed across different servers or cloud environments.
    - In the GxP field, the ability to uniquely identify the inference environment is critically important from the perspectives of traceability and data integrity. By managing container images under version control and recording them in association with specific AI model outputs, it becomes possible to retrospectively verify "which version of the model was executed in which environment," thereby enhancing explainability for audits and re-analysis.
- Verification of Performance Requirements
  - Inference speed (latency) requirements
  - Throughput (concurrent processing capability)
  - Resource usage (CPU, GPU, memory)
- Security and Access Control
  - Prevention of unauthorized access to the AI model
  - Confidentiality protection of Input data
  - Recording of audit trails
- Fail-safe Functions
  - Alternative processing upon model inference failure
  - Escalation when the Confidence Score falls below a threshold
  - Warning upon detection of abnormal Input

GAMP 5 2nd Edition emphasizes the shift from the traditional V-model to a risk-based testing approach, and terms such as IQ/OQ/PQ are not mandatory. However, since the IQ/OQ/PQ classification is still widely used in CSV practice, this document organizes them as follows for ease of understanding: Installation verification of the AI Subsystem corresponds to IQ, verification of functions and connectivity corresponds to OQ, and verification of end-to-end business processes integrated with the Core System corresponds to PQ (User Acceptance Testing: UAT). However, the names and scope of test activities should be defined according to each company's CSV procedures.

The GAMP 5 2nd Edition category for an AI Subsystem depends on how it was acquired. Custom-developed AI Subsystems typically fall under Category 5 (Custom Software). Those configured on commercial platforms fall under Category 4 (Configured Software), and commercial AI services with limited configuration options may correspond to Category 3 (Non-Configured Software). Each company should establish the appropriate classification within its QMS, based on the platform type, configuration approach, and model risk evaluation results.

5.8 Step 10: Validation of the Core System

[Discussion]

Similar to the previous section, Step 10 has been added based on GAMP 5 2nd Edition. This step corresponds to the CSV activities for the Core System (GxP System), which is responsible for collecting Input and providing the Output to the end-user. The fundamental approach to quality assurance should follow the CSV and Computer Software Assurance (CSA) procedures defined by each company.

Regarding activities related to the AI system, verification must cover not only the User Acceptance Testing (UAT) of the Core System as a standalone unit but also the entire end-to-end process, from the provision of Input information to the utilization of the Output. Particular attention should be paid to the following points:

- The expected Output is correctly received and displayed in response to the Input.
- The business process for utilizing the Output is appropriate.

In the CSV of the GxP Core System interacting with the AI Subsystem, the following practical considerations are included:

- Reliability of Data Pipelines
  - Automated process for Input data collection
  - Reproducibility of data preprocessing
  - Data quality check functions
- Integration Testing with the AI Subsystem
  - Normal case: E2E testing within the assumed Input range
  - Abnormal case: Processing of unexpected Input (missing values, out of range, outliers)
  - Boundary value: Confirmation of HITL operation in areas where the model is uncertain
- Audit Trails
  - Which model version performed the inference
  - Linking Input data with model inference results
  - Records of approval/rejection by HITL

5.9 Step 11: Operation and Maintenance

[Overview]

Step 11 refers to the entire operational phase of the AI model, following Steps 0 through 10, which includes the Proof of Concept (PoC). This document organizes Step 11 into two phases: the intensive launch period immediately after release, known as "Hypercare," and the subsequent phase of "steady operation (life cycle maintenance)".

This section outlines "B. Special Consideration: Life Cycle Maintenance of the Credibility of AI Model Outputs in Certain Contexts of Use" in the FDA AI Guidance.

In the FDA AI Guidance, "life cycle maintenance" refers to a set of planned management activities designed to ensure the AI model remains "fit for use" throughout the drug product life cycle. AI models are data-driven and possess the characteristic of performance variation due to redevelopment. In particular, for Adaptive AI models that continuously learn from new data during operation, it is essential that their outputs consistently maintain a certain level of performance.

To achieve this, model performance metrics must be continuously monitored using a risk-based approach, and both planned and unplanned changes must be managed within the PQS. For example, the impact of changes in the manufacturing process on AI model performance should be evaluated, and model re-training or re-testing should be conducted depending on the degree of impact. Significant changes that affect performance must be reported to regulatory authorities in accordance with applicable regulatory requirements.

Detailed plans regarding life cycle maintenance (e.g., monitoring frequency, triggers for re-testing) should be documented as part of the PQS and maintained in a reviewable state. Furthermore, the use of tools such as "Established Conditions" from ICH Q12 is highly recommended. By defining model-related elements as Established Conditions and including a change management plan in the marketing application, sponsors can obtain the regulatory authority's view in advance regarding which changes would not require prior approval. This approach enhances regulatory predictability and enables efficient model maintenance and management.

[Discussion]

The FDA AI Guidance addresses the operation phase in section "B. Special Consideration: Life Cycle Maintenance of the Credibility of AI Model Outputs in Certain Contexts of Use." Incorporating these considerations, this document defines Step 11 to include both "Hypercare" (to address the initial stabilization period immediately after release) and "Operation" (for full-scale operation). This section describes the regulatory background, inherent risks, and management methodologies regarding "Special Considerations" for the activities in these Steps.

Establishing an AI model's credibility requires assurance throughout its life cycle, similar to traditional CSV activities. However, compared to traditional static computerized systems, the scope handled by AI models has expanded in the following ways:

- **Temporal expansion:** Long-term utilization of AI.
- **Data expansion:** Handling of daily dynamic data, as opposed to static data (e.g., fixed datasets).
- **Role expansion:** Shift from generating a one-time report to continuous monitoring and ongoing support of operator judgment.

Accompanying this expansion of requirements, Credibility activities should also move toward a more dynamic management approach.

AI models, especially those used continuously in manufacturing processes, carry the risk of "model drift," where performance degrades over time due to a divergence between the training data and the data encountered during actual operation. This performance degradation can manifest, for example, as false negatives (overlooking defective products) in visual inspection, posing a direct hazard to product quality, patient safety, and data reliability. Therefore, beyond evaluating conformity to the COU in the initial 7 Steps, a life cycle approach that continuously monitors and maintains model performance is essential.

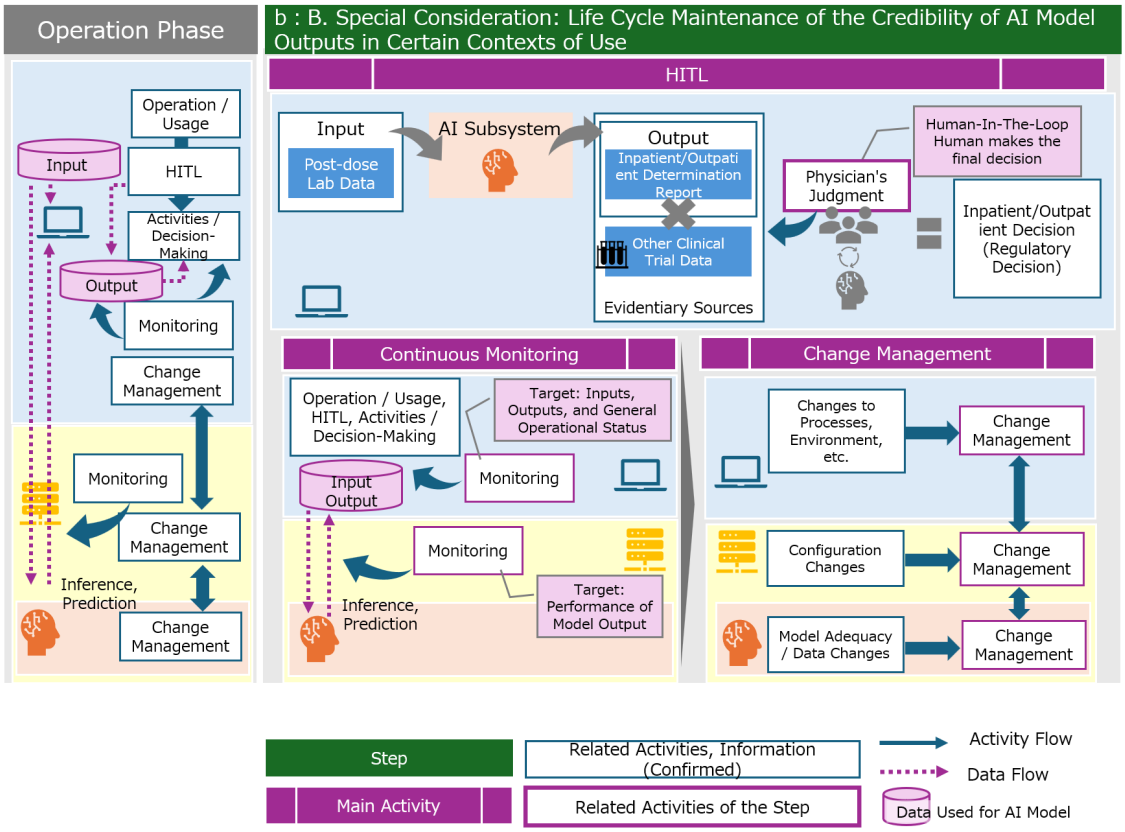

Figure 20: Special Considerations for the Operation Phase

The main management methods are detailed below.

- HITL
  - The purpose of HITL is to mitigate the risk of decision-making based on the AI model's Output. Typical HITL activities can be categorized into two types:
    - **Human Review of AI Output:** Humans review the Output generated by the AI model. Depending on the criticality of the process and the model's test level, this may involve reviewing all Outputs or reviewing Outputs based on pre-defined criteria.
    - **Utilization of Parallel Activity Results:** This involves using results from activities conducted in parallel with, but separate from, the AI model. An example of this is the "independent verification performed by personnel on a sampling basis" described in the GMP example in Section 5.2 Table 4.
  - In HITL scenarios where humans make the final decision using the AI model's Output, particularly where HITL is incorporated into the COU to reduce the scope of model testing, records of this human decision process must be retained.

- During Hypercare, if risks are detected (e.g., HITL is not functioning appropriately), reviewing and revising the operational process should be considered.

- Continuous Monitoring
  - A systematic monitoring plan covering model performance metrics (Output), statistical characteristics of input data (Input), and the actual usage environment/conditions must be formulated and executed. This enables early detection of model drift and the initiation of Corrective and Preventive Actions (CAPA).
  - Performance metrics should be defined to monitor whether Input data remains within the range of the training data at the time of model development (confirmation of Input Sample Space) and consistent with performance metrics. Typical statistical monitoring methods include PSI (Population Stability Index) to quantify changes in input feature distribution, time-series comparison of output score distributions (e.g., KS test), and tracking of confusion matrix trends. Regardless of the method adopted, sponsors are encouraged to pre-define alert thresholds and their rationale in the monitoring plan, as well as escalation procedures when thresholds are exceeded (e.g., initiation of investigation, consideration of retraining, temporary suspension of operation).
  - During Hypercare, user feedback is also a critical monitoring item. For example, confirming user comments such as "the screen display is difficult to understand" or "it is confusing to judge specific values" is beneficial for the appropriate and effective operation of the AI system, leading to improvements such as changes in screen display or additional training.
  - Furthermore, in the operational phase of the AI model, pharmaceutical companies are encouraged not only to individually verify technical monitoring results such as performance evaluation metrics and data drift indicators but also to establish regular meetings for relevant departments, including business users, IT, data scientists, and QA or CSV personnel, to share information. In line with the concept of "Multidisciplinary expertise" outlined in the "Guiding Principles of Good AI Practice in Drug Development" (hereafter referred to as AI Principles) issued jointly by the FDA and EMA in January 2026, continuous discussion and consensus-building among stakeholders regarding the interpretation of model outputs, impacts on business processes, operational challenges, and future improvement strategies will facilitate the maintenance of the AI model's credibility and GxP compliance throughout its life cycle.
- Change Management
  - All events that impact the validity of the model, such as model re-training or changes to processes/IT infrastructure that may impact performance, must be evaluated, approved, and documented under established change management procedures.

- Any change to the model, the system, or the process in which it is used (including changes to physical objects used as input) must be evaluated to determine the necessity of model re-testing. Tested models must be subject to configuration management before release, and methods to detect unauthorized changes must be established. Annex 22 currently does not allow autonomous changes to AI models, such as self-learning updates.

- Model Card (Information Disclosure to End-Users)
  - While the human role is critical in HITL, it is essential that end-users understand the AI's characteristics before use. In the "Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations" (FDA AI Medical Device Guidance, Jan 2025), the FDA recommends disclosing "what kind of AI model it is" to end-users via an easy-to-understand "Model Card." This concept is considered applicable to the pharmaceutical GxP area as well. The "Model Card" aims to help end-users understand the model's overview, performance, bias, prohibitions ("what not to do"), and limitations, ensuring safe and effective use without over-reliance. It is recommended to provide the following information in manuals and training, referencing these items:
    - **Intended Use:** Who are the intended users? What is the specific intended use?
    - **Performance:** What level of accuracy (Sensitivity, Specificity, etc.) was demonstrated in testing?
    - **Limitations and Bias:** Under what situations or for which subjects does performance degrade? (e.g., "Potential to overlook specific rare cases," "Verification in Asian populations is insufficient.")
    - **Usage:** What input data should be used? How should results be interpreted for decision-making?
    - **Risks:** What are the risks related to the model, data, and results? (e.g., Personal information acquisition, cybersecurity risks.)
    - **Change History:** What are the versions, history, and content of changes? (e.g., "Target age changed from 18+ to 22+." "Compatible 12-lead ECGs increased from Models A/B to A/B/C/D.")

Furthermore, this guidance suggests applying concepts from ICH Q12, such as "Established Conditions" and "Post-Approval Change Management Plans (PACMP)," to AI model management. This advanced approach involves defining the model's acceptable operational range and change management procedures in advance to seek consensus with regulatory authorities. Strategic utilization of this approach allows sponsors to increase regulatory predictability and rapidly adopt technological innovations while reducing the burden of post-approval changes.

Overall, regulatory authorities position AI models not merely as IT tools but as critical process components that directly impact patient safety, product quality, and data reliability. The key to compliance and project success lies in seamlessly integrating AI model life cycle management into existing drug product life cycle processes and permanently demonstrating credibility through continuous monitoring and rigorous change management.

While the retirement phase is outside the scope of this document, GxP regulations require sponsors to define the AI model archiving policy, training data retention period, and retirement procedures within the QMS. Each company's existing SOPs for computerized system decommissioning should be referenced and extended to cover AI models and AI subsystems.

6. Regulatory Gap Analysis: FDA AI Guidance and Annex 22

Up to this point, this document has outlined the AI model life cycle from conception to operation, primarily based on FDA AI Guidance while incorporating considerations from Annex 22 (EU GMP). This section presents a gap analysis between these regulatory frameworks and proposes a strategic approach for global implementation.

The requirements presented by the FDA AI Guidance and Annex 22 differ in their primary focus. The FDA AI Guidance emphasizes the life cycle process, whereas Annex 22 places significant weight on data governance. The requirements of Annex 22 appear more stringent regarding data controls, reflecting the necessity to comply with robust European legal frameworks such as the General Data Protection Regulation (GDPR) and the EU AI Act.

In practice, the "Risk-Based Credibility Assessment Framework" (7 Steps) proposed by the FDA AI Guidance aligns well with traditional validation concepts and Computer Software Assurance (CSA), providing a useful structural "skeleton" for the development life cycle. Conversely, Annex 22 serves as a practical reference for concrete activities, particularly regarding specific controls such as "Test Data Independency." Therefore, these regulations should not be viewed as conflicting, but rather as complementary. It is essential to understand both perspectives deeply and integrate them into corporate activities.

6.1 Comparison of FDA AI Guidance and Annex 22

The gaps between the regulations are compared.

Table 11: Gaps and Considerations between FDA AI Guidance and Annex 22

Item	FDA AI Guidance	Annex 22	Gap & Considerations
Overall Framework	Proposes a conceptual "Risk-Based Credibility	Provides specific technical requirements	FDA AI Guidance focuses on "How to

Item	FDA AI Guidance	Annex 22	Gap & Considerations
	Assessment Framework" (7-Step).	supplementing Annex 11. More prescriptive.	think" (Strategy), while Annex 22 focuses on "What to do" (Tactics). A comprehensive approach can be achieved by combining both perspectives.
Scope	Assumes broad AI use across the entire drug product life cycle.	Limited to "Static" and "Deterministic" models. Dynamic models and Generative AI are currently out of scope for critical applications.	Annex 22 currently has a narrower scope, likely inferring the initial stage of AI use in GxP. Caution is required when using dynamic models in Europe.
Data	Emphasizes the importance of "Fit for Use" (data relevant to the purpose).	Strictly requires ensuring technical and procedural independence of Test Data.	Annex 22 requires more specific and stringent controls regarding data governance. Management of Test Data independence is a critical compliance issue.
Model Evaluation	Requires model performance evaluation including "Confidence Intervals*" and description of limitations.	Specifically requires evaluation axes for individual AI outputs, such as "Explainability" and "Confidence Score**."	FDA AI Guidance requires reliability as a system/procedure, while Annex 22 requires reliability of individual results. Approval of "Black Box" models may be challenging without explainability measures.
Operation	Points out the importance of maintenance	Requires specific monitoring items such as performance	Continuous verification activities in the operational phase are mandatory.

Item	FDA AI Guidance	Annex 22	Gap & Considerations
	throughout the life cycle.	monitoring and input data monitoring.	
Data Integrity	Applies data integrity requirements for computerized systems in general, such as audit trails, access controls, and ensuring the integrity and authenticity of records, based on existing GxP guidance.	Explicitly requires integrity and traceability of training/test data and AI outputs in Annex 22 and related documents. Emphasizes ensuring data integrity throughout the AI lifecycle, including technical measures such as data fixing via hash values (tamper-proofing) in addition to dataset version control and change history management.	There is a clear distinction in that FDA AI Guidance emphasizes a "General Framework" approach by applying existing GxP requirements, whereas Annex 22 delves into "Specific Technical Details" regarding AI-unique data processing, such as hashing and dataset management requirements.

* Confidence intervals: A performance evaluation metric for the AI model as a whole.

** Confidence Score: The degree of certainty for each individual output generated by the AI model.

Annex 22 limits the scope of the guideline to "static models" where performance does not change by incorporating new data during operation. Annex 22 explicitly states that dynamic/adaptive AI models, Generative AI, and LLMs (probabilistic output models) **should not be used** for critical GMP applications that directly impact patient safety, product quality, or data integrity.

On the other hand, the FDA AI Guidance presents a risk-based framework to maintain "Credibility" throughout the AI model life cycle. It does not explicitly prohibit Adaptive/Dynamic models like Annex 22 does. Rather, it implies a stance where model updates and retraining are anticipated, premised on life cycle maintenance including change management and drift monitoring. However, when using such AI models for regulatory decision-making, "early consultation with the FDA" is strongly recommended.

In this document, based on the differences in stance between FDA AI Guidance and Annex 22, we recommend the following concepts as a basic policy for global deployment:

- For critical GMP applications, the principle should be to adopt "static and deterministic models" that meet Annex 22 requirements as a common global baseline regardless of the region. In other words, even in the US, for uses directly linked to regulatory decisions, static models that do not perform automated learning during operation should be used. Model revisions should be handled as new versions via offline retraining, Credibility Assessment, and change management.
- Even in the US, using self-learning/continuous learning Adaptive AI models for regulatory decision-making should not be interpreted as automatically "allowed." When considering the use of such models, sponsors are encouraged to clarify the COU and model risk and then prepare a "pre-defined change management plan" like the Predetermined Change Control Plan (PCCP) for AI in the medical device field. Specifically, sponsors will need to document in advance the scope of assumed model changes (retraining, algorithm improvement, etc.) and the life cycle maintenance plan detailing the verification, impact assessment, and release processes required in such cases.
- Furthermore, regarding the specific plan, COU, and model risk, we recommend conducting early consultation (prior engagement) with regulatory authorities for each individual case to confirm the applicability and operational conditions of Adaptive/Dynamic AI models in a form that aligns with the authority's expectations. This facilitates exploring a "practical landing point" where, for example, "static models are the minimum requirement for Annex 22, while limited adaptiveness is recognized within a pre-agreed framework for the US."
- For non-critical uses (e.g., operational efficiency or supplementary analysis), if Adaptive/Generative AI is used, it is desirable to position them within the QMS as "uses logically and technically separated from critical GMP processes" and confirm consistency with Annex 22 and local guidelines.

As described above, adopting the static model framework required by Annex 22 as a global requirement, while considering additional advanced Adaptive AI in the US, enables AI model design and operation consistent with US and European expectations. This assumes a "pre-defined change management plan similar to PCCP" and prior consultation with authorities.

Furthermore, a shared tenet between the FDA and EMA is the necessity of a cooperative structure with experts. As reiterated throughout this document, AI model development requires that Subject Matter Experts ("SMEs") across various fields (e.g., data scientists) be effectively integrated into the project and reach consensus. The "Guiding Principles of Good AI Practice in Drug Development"

("AI Principles"), issued jointly by the FDA and EMA in January 2026, emphasize "5. Multidisciplinary expertise" and "10. Clear, essential information." Since the AI Principles advocate for the "use of experts in each domain" and collaboration among SMEs with different backgrounds, the AI Principles call for "using language that can be clearly understood by all parties." In AI model development, terms taken for granted in existing frameworks may not apply, or there may be a lack of background understanding among personnel. Therefore, paying attention to communication is extremely important.

6.2 Regulatory Trends in Japan

At present, the Ministry of Health, Labour and Welfare (MHLW) and the Pharmaceuticals and Medical Devices Agency (PMDA) have not issued guidance corresponding to the "Credibility Assessment" of AI models in drug development. However, the "AI Guidelines for Business" (March 2025) by the METI, referenced in this document, serves as a general AI governance guideline in Japan.

Additionally, through prior consultation with the PMDA, pharmaceutical companies can agree on the management and use of each AI system, enabling practical responses referencing the FDA AI Guidance and Annex 22. Japanese pharmaceutical companies are recommended to build a system capable of responding to future domestic regulations by satisfying the requirements of the FDA and EMA.

For reference, relevant domestic notifications are listed below:

- Regarding the Relationship between the Use of Programs Supporting Diagnosis and Treatment using Artificial Intelligence (AI) and the Provisions of Article 17 of the Medical Practitioners Act (MHLW, December 2018):
  - This notification clarifies that even when using AI for medical support, the ultimate judgment and responsibility lie with the physician, and such acts are positioned as medical practice under Article 17 of the Medical Practitioners Act.
- Guideline for the Utilization of Medical Digital Data for AI Research and Development (MHLW, March 2024):
  - The guideline systematizes the legal basis and specific processing/operational procedures for smoothly utilizing data accumulated in medical institutions as pseudonymized information for AI product development by private companies.

7. Limitations and Application of This Document

While this document serves as a useful guide for understanding the regulations and basic practical principles regarding the Credibility Assessment Framework for AI models, AI Users and practitioners should be aware of its limitations. The primary references, the FDA AI Guidance and Annex 22, are drafts issued in 2025 and may not fully reflect the rapidly evolving technological and regulatory landscape. For instance, although this document excludes detailed explanations of the Retirement Phase based on current draft guidance (see Section 5.9), requirements for the retirement and archiving of AI systems and related data are likely to become critical issues from a GxP lifecycle management perspective.

Furthermore, for pre-trained models such as Foundation Models and LLMs, the evaluation methods outlined in the FDA AI Guidance remain conceptual at this stage and have limited applicability as practical guidelines.

Technology companies providing pre-trained models publish Model Cards and Technical Reports on model performance, limitations, and safety assessments. However, the following information required for use in a GxP environment may not be sufficiently disclosed:

- Detailed configuration of training data (sources, sampling methods)
- Specific procedures for data pre-processing
- Quality control processes for annotation
- Details of hyperparameters used during training

This lack of transparency creates a gap between available information and the pharmaceutical industry's responsibility to prioritize patient safety. Additionally, traditional QMS-centric supplier audits often lack the perspective of training data management, necessitating a new, industry-wide approach.

This places a significant responsibility on pharmaceutical companies to tailor Credibility Assessment activities to their specific COU, similar to the expectations for in-house AI model development. The most critical concern is the potential erosion of "Critical Thinking," a key control measure in GxP, due to complacency or superficial judgments, such as viewing AI merely as an "auxiliary tool." Critical thinking is the intellectual process of actively questioning the reliability of AI Output ("Why can this output be trusted?") and assessing potential risks ("What if the output is incorrect?"), rather than accepting it unconditionally. The omission of this process leads directly to logical judgment becoming a mere formality.

Therefore, this document should not be used merely as a checklist, but rather as a conceptual framework to foster and apply critical thinking. Ultimately, the most robust Credibility Assessment lies not in the document or system itself, but in the critical thinking exercised by each individual.

8. Conclusion

This document aims to assist a wide range of stakeholders, including those without expertise in AI technology, in understanding the quality assurance of AI systems in alignment with current regulatory trends. The quality assurance concepts presented herein are intended to serve as a compass for realizing the vast potential of AI tools while keeping patient safety as the top priority. We hope that the discussions in this document will be translated into concrete practices within the JPMA, such as optimizing facility selection in the GCP area and refining adverse event determination and signal detection in the PV area.

During the preparation of this document, we attempted to simulate practical scenarios to define specific features for performance evaluation and potential deliverables for each step. However, due to the challenges in defining appropriate COUs and quantifying the specific benefits of AI technology at this stage, we have left these detailed simulations for future work.

In addition, due to restrictions such as the terms of use for pre-trained models, conducting supplier assessments centered on traditional QMS audits remains difficult. For this reason, public documents and third-party certifications are expected to become the primary basis for evaluation.

Therefore, the continuous discussion and accumulation of knowledge across the industry are essential. We intend to continue these activities, working alongside SMEs in each field to support the practical application of this framework. By promoting innovation that drives transformation within pharmaceutical companies in coordination with regulatory authorities, we aim to integrate credible AI systems into GxP processes and truly contribute to improving the Quality of Life (QOL) for patients.

9. References

- Considerations for the Use of Artificial Intelligence to Support Regulatory Decision-Making for Drug and Biological Products Guidance for Industry and Other Interested Parties (FDA, January 2025, Draft)
- EU GMP Annex 22: Artificial Intelligence (EC, July 2025, Draft)
- AI Guidelines for Business (METI, March 2025, Ver 1.1)
- GAMP 5 - A Risk-Based Approach to Compliant GxP Computerized Systems (ISPE, July 2022, 2nd Edition)
- Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations (FDA, January 2025, Draft)
- Computer Software Assurance for Production and Quality System Software Guidance for Industry and Food and Drug Administration Staff (FDA, February 2026)
- Guiding Principles of Good AI Practice in Drug Development (FDA, EMA, January 2026)
- ICH Q12 Technical and regulatory considerations for pharmaceutical product lifecycle management - Scientific guideline (ICH, November 2019)
- Regarding the Relationship between the Use of Programs Supporting Diagnosis and Treatment using Artificial Intelligence (AI) and the Provisions of Article 17 of the Medical Practitioners Act (Ministry of Health, Labour and Welfare, December 2018)
- Guideline for the Utilization of Medical Digital Data for AI Research and Development (Ministry of Health, Labour and Welfare, March 2024)
- Overview of the FDA CSA Draft Guidance and Examination/Consideration of its Application to the GxP Area (Japan Pharmaceutical Manufacturers Association (JPMA), Drug Evaluation Committee, Electronic Data Management Committee, FY2023 Task Force 4, July 2024)

Authors

Electronic Data Management Committee

Chairperson	Naoki Sakuma	Teijin Pharma Limited
Vice Chairperson	Manabu Inoue	MSD K.K.
	Miki Someya	Pfizer R&D Japan G.K.
	Hiroshi Watanabe	DAIICHI SANKYO COMPANY, LIMITED

Electronic Data Management Committee Task Force 4 Sub-task Force Members (in no particular order)

TF Leader	Hiroshi Watanabe	DAIICHI SANKYO COMPANY, LIMITED
STF Leader	Tetsuro Miyake	Bayer Yakuhin, Ltd.
	Susumu Nakao	Eisai Co., Ltd.
	Kohei Shinkai	Kissei Pharmaceutical Co., Ltd.
	Yuki Nakamura	Taisho Pharmaceutical Co., Ltd.
	Yoko Nakajima	Chugai Pharmaceutical Co., Ltd.
	Hiroshi Watanabe	DAIICHI SANKYO COMPANY, LIMITED