IMDRF International Medical Device Regulators Forum

1
2
3

Draft

IMDRF/AIML WG/N93 DRAFT: 202X

Technical Framework for Artificial Intelligence Life Cycle Management

AUTHORING GROUP

Artificial Intelligence/Machine Learning-enabled Working Group

4
5

7 April 2026

Preface

[Name], IMDRF Chair

Contents

1. Introduction

The International Medical Device Regulators Forum (IMDRF) recognizes the growing role of artificial intelligence (AI) technologies in medical devices and has continued its efforts to establish a harmonized approach for the oversight, evaluation, and use of AI-enabled medical devices. Building on its previous work, including the publication IMDRF/AIML WG (Artificial Intelligence/Machine Learning-enabled Working Group)/N88 FINAL:2025: *Good Machine Learning Practice (GMLP) for Medical Device Development: Guiding Principles*, IMDRF is introducing this document focused on considerations for AI-enabled medical devices across their total product life cycle.

The GMLP principles describe foundational best practices for the development of AI-enabled medical devices, emphasizing areas such as data quality, model transparency, performance evaluation, and the role of multidisciplinary expertise. These principles underpin each step of the AI life cycle, and this document provides relevant GMLP references to help provide a foundational understanding of applicable principles.

Together with the broader IMDRF mission, this work contributes to the establishment of globally harmonized considerations that, when applied across a device's total product life cycle, can foster innovation while protecting public health.

2. Purpose and Scope

2.1. Purpose of the document

This document provides an internationally harmonized technical framework to help promote responsible innovation and patient-centricity by facilitating the secure, safe, ethical, and effective design, development, deployment, maintenance, use, and decommission of AI-enabled medical devices. .

The purpose of this document is to:

- Provide foundational information on AI-enabled medical device life cycle management;
- Highlight universal concepts applicable to all life cycle steps;
- Provide an overview of concepts and considerations for each step of the AI-enabled medical device life cycle; and
- Provide references to applicable, internationally-recognized standards and resources.

The document is not meant to:

- Replace or conflict with existing IMDRF publications such as those published by the Software as a Medical Device or Cybersecurity Working Groups (WG.H). However, it may in part relate to or overlap with those publications and is intended to be complementary in those circumstances. Where relevant, this document references other IMDRF publications on related topics (e.g. Risk Management);
- Serve as regulation nor guidance. It is not intended to be an interpretation of any jurisdiction's laws and regulations, does not provide recommendations for market submission within a specific jurisdiction, and does not imply a convergence of regulations across jurisdictions. Instead, this document aims to describe harmonized concepts and general considerations for the AI-enabled medical device life cycle. Individual jurisdictions may apply and align some or all of these concepts to their particular regulatory framework; and
- Provide specific information on how to prepare a regulatory submission. However, in some of its sections, documentation is identified that may be helpful when communicating with regulatory authorities.

2.2. Scope of the document

This document is intended for manufacturers that are developing AI-enabled medical devices to identify considerations as they make critical technical and governance choices. It is intended to provide considerations specific to where AI-enabled medical devices may necessitate new or different approaches than those more broadly applicable to medical devices and medical device software, building upon existing work and harmonization efforts IMDRF has published. Rather than exhaustively cover all steps of the AI-enabled medical device life cycle, the document highlights key considerations as well as internationally-recognized standards and resources for manufacturers to additionally consider. Given the rapid pace of advancement in the field of AI, this document attempts to reflect the most current considerations for manufacturers of AI-enabled medical devices.

This document applies to AI-enabled medical devices, which include Machine Learning-enabled Medical Devices (MLMD). MLMD are medical devices that use machine learning, in part or in whole, to achieve their intended purpose[1]. Machine Learning (ML) models are developed by ML training algorithms through analysis of data, without models being explicitly programmed[1]. While this document is intended to apply generally to MLMD, certain types of MLMD that are enabled by or that incorporate generative AI, autonomous or adaptive models may warrant additional considerations. While this document makes some recommendations for these subsets of MLMD, some aspects of this AI life cycle may differ for such technologies and this document is not intended to be comprehensive.

The concepts of a "model" and "AI-enabled medical device" appear throughout this document but are not used interchangeably. A model is a "mathematical construct that generates an inference or prediction based on new input data, and is the result of an ML training algorithm learning from data".[2] As stated above, AI-enabled medical devices, that include MLMD, are devices that use AI or ML, in part or in whole, to achieve their intended medical purpose, and incorporate one or more "models" into their design to achieve this purpose. Certain steps of the life cycle described in this document focus on the model (e.g., Data Collection and Management and Model Building and Tuning), while other steps describe considerations for AI-enabled medical devices that incorporates the model(s)). This distinction will be highlighted in the sections throughout this document.

Explainability and Interpretability are also discussed throughout this document. Explainability "refers to a representation of mechanisms underlying AI systems' operation"[3]. It includes not only understanding how decisions are made, but also why they are appropriate and trustworthy in the clinical context. Interpretability "refers to the meaning of AI systems' output in the context of their designed functional purposes"[3]. In other words, the ability for clinicians and healthcare professionals, patients, and other users to comprehend how the device arrives at its outputs or recommendations. Understanding the difference between these concepts is important when applying considerations outlined throughout this document.

This document acknowledges that AI-enabled medical devices have a variety of users and can be deployed in a variety of environments, including at medical institutions, hospitals, or in other healthcare settings (collectively referred to as "sites") as well as on a patient's own consumer electronics or in a patient's home. While the considerations in this document are intended to apply across the spectrum of intended use environments, some considerations are specific to the deployed environment and are indicated as such throughout the document.

[1] IMDRF/AIMD WG/N67 (Edition 1):*2022 Machine Learning-enabled Medical Devices: Key Terms and Definitions*

[2] IMDRF/AIMD WG/N67 (Edition 1):2022 *Machine Learning-enabled Medical Devices: Key Terms and Definitions*

[3] National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.

3. References

IMDRF/AIML WG/N88 FINAL:2025: *Good Machine Learning Practice (GMLP) for Medical Device Development: Guiding Principles*

IMDRF/AIMD WG/N67 (Edition 1):*2022 Machine Learning-enabled Medical Devices: Key Terms and Definitions*

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf.

ISO 13485 Third Edition 2016-03-01 *Medical Devices—Quality Management Systems—Requirements for Regulatory Purposes*

IMDRF/SaMD WG/N23 FINAL:2015 *Software as a Medical Device (SaMD): Application of Quality Management System*

ANSI/AAMI/ISO 14971:2019 *Medical devices—Application of risk management to medical devices*

ISO/TR24971:2020 *Medical devices — Guidance on the application of ISO 14971*

AAMI TIR 34971:2023 *Application of ISO 14971 to machine learning in artificial intelligence— Guide*

ENISA. (2021, December 14). *Securing Machine Learning Algorithms*, Enisa Europa. https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms

Apostol Vassilev (NIST), Alina Oprea (Northeastern University), Alie Fordyce (Robust Intelligence), Hyrum Anderson (Robust Intelligence) (2024, January) NIST AI 100-2 E2023 *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. NIST. https://csrc.nist.gov/pubs/ai/100/2/e2023/final

IEC 62304:2006/A1:2016 *Medical device software – Software life cycle processes*

IMDRF/MC/N79 DRAFT:2023 Guiding Principles to Support Medical Device Health Equity

ISO/IEC 5259-4:2024 *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework*

IMDRF/CYBER WG/N60 FINAL:2020 *Principles and Practices for Medical Device Cybersecurity*

IMDRF/CYBER WG/N70 FINAL:2023 (Edition 1) *Principles and Practices for the Cybersecurity of Legacy Medical Devices*

IMDRF/CYBER WG/N73 FINAL:2023 (Edition 1) *Principles and Practices for Software Bill of Materials for Medical Device Cybersecurity*

IMDRF/SaMD WG/N41 FINALFINALN41 FINAL:2017 *Software as a Medical Device (SaMD): Clinical Evaluation*

ISO/IEC 23053:2022 *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

ISO/IEC TS 25058:2024 *Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guidance for quality evaluation of artificial intelligence (AI) systems*

ISO/IEC 25059:2023 *Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems*

IEEE 2941-2021 IEEE Standard for Artificial Intelligence (AI) Model Representation, Compression, Distribution, and Management

IMDRF/UDI WG/N48 FINAL:2019 *Unique Device Identification system (UDI system) Application Guide*

185 IMDRF/GRRP WG/N52 FINAL:2024 (Edition 2) *Principles of Labelling for Medical Devices and IVD*
186 *Medical Devices*

187 IMDRF/SaMD WG/N81 DRAFT:2024 *Medical Device Software: Considerations for Device and Risk*
188 *Characterization*
189
190 ISO 14155:2020 *Clinical investigation of medical devices for human subjects - Good clinical practice*

191 ISO/IEC 62366-1:2015 *Medical devices Part 1: Application of usability engineering to medical devices*

192

4. Universal Concepts that apply across the AI Life Cycle

This section highlights universal concepts applicable to all steps of an AI-enabled medical device's life cycle. Many of these universal concepts and their considerations are not new or unique for AI-enabled medical devices, but can be adapted for the unique characteristics of these devices.

4.1. Quality Management System (QMS)

A QMS is a comprehensive framework of policies, processes, and procedures that medical device manufacturers implement to ensure consistent quality throughout the device life cycle. It is important for manufacturers to implement a QMS that complies with applicable requirements and established standards as defined by regulatory authorities in their regulations. This may include, for example, the international standard ISO 13485 Third Edition 2016-03-01 *Medical Devices—Quality Management Systems— Requirements for Regulatory Purposes* (hereafter referred to as ISO 13485:2016). Furthermore, IMDRF/SaMD WG/N23 FINAL:2015 *Software as a Medical Device (SaMD): Application of Quality Management System* (hereafter referred to as IMDRF/SaMD WG/N23 FINAL:2015) provides guidance to manufacturers and regulators on QMS practices specific to SaMD, which can generally be applied to AI-enabled medical devices.

AI-enabled medical devices benefit from implementation of scalable life cycle support processes that emphasize safety-focused risk management throughout all life cycle steps. For example, QMS requirements management captures functional specifications as well as clinical environment considerations, such as how AI outputs will be interpreted by healthcare providers or patients. Additionally, configuration management and control processes maintain traceability of training data and AI model versions, which is especially important for models that adapt to new data in their intended use environment. Finally, comprehensive post-market monitoring and surveillance mechanisms can help actively monitor performance of the AI-enabled medical device in real-world environments and enable rapid response to emerging safety concerns or performance degradation over time.

4.2. Risk Management

Risk management principles and processes for AI-enabled medical devices follow the same fundamental framework established for other medical devices, including SaMD. As with all medical devices, AI-enabled medical devices carry risks that are necessary for manufacturers to systematically address. Manufacturers are responsible for evaluating these risks through comprehensive risk analysis and assessment processes, implementing appropriate controls to reduce identified risks to acceptable levels, ensuring medical devices do not pose unacceptable risk to patients, users, or others, and demonstrating that clinical benefits outweigh residual risks through appropriate evidence and documentation.

This section highlights key risks that are unique to AI-enabled medical devices and that warrant careful consideration and mitigation throughout the life cycle. The examples provided below serve as starting points for AI-related risk management activities and do not constitute an exhaustive list of all possible AI-related risks that a medical device may encounter.

4.2.1. Risks Related to Information

The “black box” nature of some AI models makes it challenging to understand how and why certain outputs are produced or why certain decisions are made by the model (when a model is embedded and controls hardware or autonomously makes decisions), creating unique transparency, explainability, interpretability concerns. Risks related to information can include:

- Inaccurate or misleading outputs
- Incomplete information presentation, including lack of explanation on probability of error

4.2.2. Risks Related to the Human-AI Interaction

The integration of AI-enabled medical devices into the clinical workflow can significantly alter interaction dynamics in healthcare settings. One such manifestation of this is automation bias, which refers to a tendency of users to rely on automated systems over their own knowledge or skills when making decisions. AI-enabled medical devices may also be perceived as having human-like cognitive or reasoning abilities, potentially leading to over-reliance that compromises independent reasoning and decision-making. These types of risk may develop over time and may not be detectible in testing or early use of a device. Risks related to the human-AI interaction can include:

- Over-reliance or automation bias
- Under-reliance or dismissal
- Workflow disruption
- Verification fatigue
- Information overload (e.g., alert fatigue)
- De-learning of clinical knowledge (due to over-reliance over time)

4.2.3. Risks Related to Model Training and Data Quality

AI-enabled medical devices may be uniquely vulnerable to performance degradation from, for example, data drift when they are exposed to real-world data distributions that significantly differ from the training data on which their model(s) were developed. Often, the substantial volume and complexity of data required for training introduces scalability challenges in data curation, annotation, and quality assurance, which limits generalizability. Risks related to model training and data quality can include:

- Training data bias (i.e., underperformance due to lack of training data for certain demographic groups)
- Data drift, including changes in the context of use, such as patient demographics
- Incomplete or missing data
- Labelling or annotation errors
- Out of distribution (OOD) inputs
- Knowledge corpora fragmentation or misalignment (e.g., discrepancies between the model’s training or reference data and current medical standards or the specific clinical environment where it will be deployed)

4.2.4. Risks Related to Deployment and Post-Market Monitoring and Performance

AI-enabled medical devices may rely on complex algorithms, substantial computational resources, and interconnected system dependencies that create novel failure modes where infrastructure inadequacies, integration issues, or deployment errors can compromise medical device performance and potentially lead to patient harm if not properly identified and controlled. Risks related to deployment and post-market monitoring and performance can include:

- Integration, interoperability, or compatibility issues
- Performance degradation due to changes in third-party, general-purpose models (see Section 5.3 for more information on these types of models) incorporated into a device
- Software governance issues including lack of version control when the AI model or device is modified
- Infrastructure inadequacy and computational or scalability constraints
- Misaligned performance and calibration if context-of-use differs from validation

Additional Standards and References to be Considered

- ANSI/AAMI/ISO 14971:2019 *Medical devices—Application of risk management to medical devices*
- ISO/TR24971:2020 *Medical devices — Guidance on the application of ISO 14971*
- AAMI TIR 34971:2023 *Application of ISO 14971 to machine learning in artificial intelligence—Guide*

4.3. Human Oversight

Human oversight, including that of clinicians, health care providers, patients, and lay users, is essential throughout the entire life cycle of AI-enabled medical devices to ensure that human and clinical expertise informs model development, validates real-world performance, and maintains appropriate human-AI collaboration that prioritizes patient safety and effective clinical decision-making.

For example, human involvement in identifying user needs for an AI-enabled medical device can help to ensure that the device design accounts for the user's ability to interpret AI outputs, override or reverse automated recommendations, and intervene or interrupt automation. As with all medical devices, usability and human factors testing with representative users can help identify potential use-related hazards or validate the design of controls for use-related hazards.

Clinician expertise and input can be leveraged in activities such as feature selection, validating the clinical relevance of models, data labelling and annotation and identifying potential biases. Clinical involvement also ensures that clinical decisions remain accurate, safe, and contextually appropriate and that complex or atypical cases are adequately understood and addressed.

Post-market monitoring and surveillance benefits from human oversight in monitoring to detect performance degradation, identify unexpected failure modes, and assess real-world effectiveness across patient populations and clinical settings. This oversight can also help manufacturers to remain aware of the possible tendency of automatically relying or over-relying on the output produced by the AI-enabled medical device (automation bias) and make adjustments for such biases.

Additional Standards and References to be Considered

- ISO/IEC 62366-1:2015 *Medical devices Part 1: Application of usability engineering to medical devices*

4.4. Cybersecurity

Similar to other medical devices, cybersecurity is an important consideration throughout the AI-enabled medical device life cycle. The large quantity of data needed for the development and validation of AI-enabled medical devices and the sensitive patient data processed during use can make these devices an attractive target for data theft. Additionally, maliciously poisoned data could negatively impact the AI-enabled medical device's performance. This section focuses on the cybersecurity related to the potential harm to an affected person, though it acknowledges that data extraction and exploitation are also possible issues.

Mechanisms to help protect access to the data used for the development and validation during the AI-enabled medical device life cycle include, but are not limited to, data anonymization, data separation, system separation, data validation, data encryption, access control, logging, auditing, model simplification, anomaly detection techniques, and periodic audits. For AI-enabled medical devices specifically, it is essential to follow best practices for mapping data sources and controlling data suppliers and labellers as part of QMS processes. This can help control unwanted bias and data poisoning stemming from outside sources in both development and post-market performance optimization (when carried out with appropriate regulatory approval).

In some instances, due to the complex nature of AI-enabled medical device design, it might be difficult to recognize when data poisoning occurs because the cause-effect relationship between input data and AI-enabled medical device output is not always transparent and explainable, such that malicious degradation may be thought to be within performance specifications of the model or mistaken for natural performance drift. In addtion to data derived vulenerabilities and complexities with transparency and explainability, the AI model itself can be under threat potentially leading to model inversion, extraction and evasion occuring. As outlined below, robust security controls and monitoring activities can help mitigate these vulnerabilities.

Using a secure product development framework to manage cybersecurity risks can help identify and reduce the number and severity of vulnerabilities in devices. Using device design processes to support secure product development and maintenance may include Threat Modeling, Cybersecurity Risk Assessments, interoperability considerations, third party software components, Cybersecurity unresolved anomalies and Risk Management.

Deployed AI-enabled medical devices are vulnerable to cybersecurity threats due to their potential reliance on continuous data flows, cloud connectivity for model updates, and complex software architectures that may create multiple entry points for malicious actors. Correct deployment and continuous post-market monitoring and surveillance, which is described in more detail in the Deployment (Section 5.5) and Operations and Monitoring (Section 5.6) life cycle steps below, is an important consideration to minimize cybersecurity threats in the post-market setting. Unwanted access may be mitigated via mechanims such as; limiting the frequency of updates to batches, hosting the AI-enabled medical device locally, if possible and ensuring security controls are centralized to manage monitoring, auditing and validation more reliabily. It is important to note that hosting of AI-enabled medical devices on their manufacturers' server may introduce additional reliability considerations that are not unique to AI-enabled devices.

Furthermore, when deploying AI-enabled medical devices that contain adaptive models that continue to learn after their initial release, it is important for manufacturers to consider how data inputs to the device in the deployed environment are controlled and validated, and who is handling and accessing the data prior to its input into the AI-enabled medical device. The primary responsibility for validation of input data resides with the legal manufacturer; in practice, additional parties such as end users and sites at which the device is being deployed may also be involved. If the model is autonomously adapting and improving, it may warrant further scrutiny of interconnected systems that deliver this data and their attack surfaces and it remains critical that updates are verified for safety prior to implementation. Again, performing post-market validation checks on algorithm performance after updates with adequate human oversight as described in Section 4.3, may help identify attempts at data poisoning or other disruptive attacks related to the AI-enabled medical device adapting to new data.

Additional Standards and References to be Considered

- ENISA. (2021, December 14). *Securing Machine Learning Algorithms*, Enisa Europa. https://www.enisa.europa.eu/publications/securing-machine-learning-algorithms

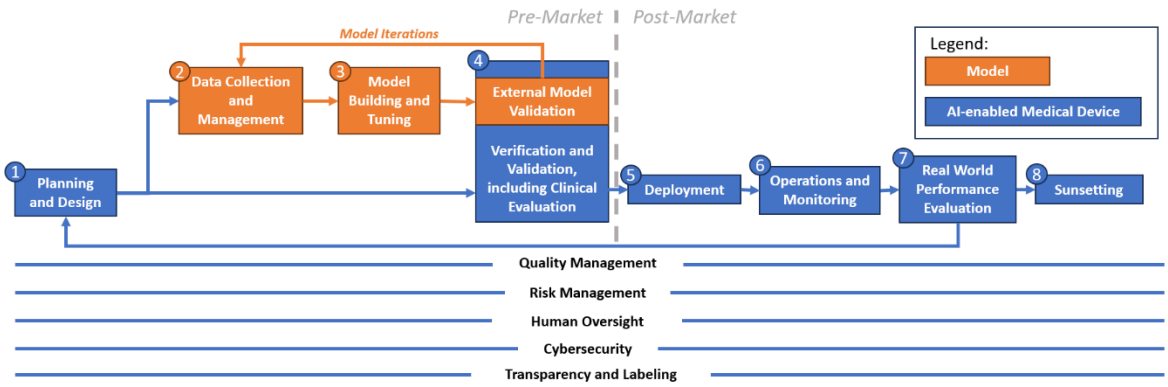
- Apostol Vassilev (NIST), Alina Oprea (Northeastern University), Alie Fordyce (Robust Intelligence), Hyrum Anderson (Robust Intelligence) (2024, January) NIST AI 100-2 E2023 *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*. NIST. https://csrc.nist.gov/pubs/ai/100/2/e2023/final

- IMDRF/CYBER WG/N60 FINAL:2020 *Principles and Practices for Medical Device Cybersecurity*

5. AI-enabled Medical Device Life Cycle Steps

The AI-enabled medical device life cycle presented in this document builds upon established frameworks, such as ISO 13485:2016, IMDRF/SaMD WG/N23 FINAL:2015, and the International Electrotechnical Commission (IEC) 62304 A1:152015 *Medical device software – Software life cycle processes* [4] standard with additional considerations for the uniqueness of incorporating AI into a medical device, including AI's data-driven nature and complexity of the human-AI interaction. The core principles of medical device and SaMD life cycles of systematic planning, requirements management, traceability, risk management and validation remain essential. The figure below depicts the steps of the life cycle described in this document, as well as the universal concepts described in Section 4 above.

Figure 1: AI-Enabled Medical Device Life Cycle



This figure represents the general life cycle outlined in this document. However, manufacturers may follow different iterations of this cycle depending on what might be needed for the model and/or device. One example of this flexibility is the pathway from Step 7 directly to Step 1 and then to Step 4. This sequence illustrates scenarios where real-world performance evaluation identifies incidents or performance drift following deployment. In such cases, a model update may not be warranted but could instead involve labelling modifications or a comprehensive re-evaluation of the model and/or device to reestablish baseline performance metrics.

[4] IEC 62304:2006 *Medical device software – Software life cycle processes* is one commonly used software development life cycle standard that can be used to develop an AI-enabled medical device software life cycle process

5.1. Planning and Design

Similar to of the planning for other medical devices, and, per IMDRF/SaMD WG/N23 FINAL:2015, "the objective of planning is to provide a roadmap to be followed during the product development life cycle" following "a methodical and rigorous plan for managing projects such as a plan-do-check-act approach". For AI-enabled medical devices, GMLP Guiding Principle 1[5] emphasizes that the intended use / intended purpose of the device is well understood, and multi-disciplinary expertise will be leveraged throughout the total product life cycle. This can be achieved with a comprehensive product definition, which includes understanding the clinical objectives, existing standards of care, use environment, potential risks and controls, relevant clinical workflows and user needs and constraints as well as team resource planning.

While the Planning and Design life cycle step encompasses many of the preliminary plans made for subsequent life cycle steps, this section generally highlights key considerations for this step, with additional details on each particular step in the sections below. Furthermore, this section points out considerations for both the AI-enabled medical device and its model(s), which both have their own unique planning and design considerations.

Model Selection and Anticipating Risks: As discussed in Section 4.2 above, the risks and implementation challenges of incorporating AI may exceed its clinical benefits for certain medical uses. It is important for manufacturers to consider whether an AI model (versus, for example, a rules-based approach) is the right tool to achieve the intended purpose of the device, and if so, what kind of model is best suited for that purpose. If the decision is made to move forward with incorporating an AI model into a medical device, selection of the appropriate model up front may streamline the risk control process; by enabling faster and more accurate identification of potential risks, manufacturers can implement preventative risk control measures rather than relying solely on corrective actions after development. This strategy is often more effective than applying risk control measures retrospectively to compensate for suboptimal model choices. Manufacturers may also want to consider opting for simplicity if explainability and interpretability is critical to the intended use environment.

Data Availability and Suitability: AI-enabled medical devices and their models rely heavily on fit-for-purpose, representative[6] data, and it is important for manufacturers to begin thinking about data availability and suitability, and generally the feasibility of data collection, during the Planning and Design step. This can help manufacturers preliminarily understand the quantity, quality and distribution of available data that would be needed for the design, development, and validation of their AI model(s). Furthermore, it is important for manufacturers to consider quantity of data needed such that training datasets can be independent from test data sets.[7]

Model, Infrastructure, and Use Requirements: In this step, it is important for manufacturers to contemplate the spectrum of technical considerations, from the architecture and model design to be employed and their requirements, such as interpretability, stability, and performance requirements, to the infrastructure needed for deployment, monitoring and maintenance.[8] For AI-enabled medical devices specifically, deployment may include site-specific localization or customization (with appropriate regulatory approval), and it is important that the infrastructure account for version control and resets or rollbacks in the event that recalls or issues occur. It is beneficial for manufacturers to not only define requirements from a technological perspective, but also for how the AI-enabled medical device and its model(s) will be used in practice, which may include additional design and labelling considerations[9] (as further described in Section 6 of this document), additional human oversight requirements, possible workflow challenges, and the security of the model and its data once deployed.

---

[5] IMDRF/AIML WG/N88 FINAL:2025: *Good Machine Learning Practice (GMLP) for Medical Device Development: Guiding Principles*

[6] See GMLP Guiding Principle 3.
[7] See GMLP Guiding Principle 4.
[8] See GMLP Guiding Principle 2.
[9] See GMLP Guiding Principle 9.

Validation and Clinical Evaluation Needs: The term validation has been used to represent different concepts within the fields of medical device development and AI model development[10]. Validation, in this context and aligned with IMDRF/AIMD WG/N67 (Edition 1):2022, means "confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use can be consistently fulfilled". In this step, it is essential to establish preliminary validation strategies and appropriate evaluation metrics (see Appendix B), including specifying acceptance criteria and thresholds for the chosen metrics as well as statistical analysis methods, which can impact how data is collected and used.

Post-Market Monitoring and Surveillance: Requirements planning for continuous monitoring of the ongoing performance of the AI-enabled medical device in the clinical workflow is crucial to ensure patient safety and device effectiveness. It is important for manufacturers to begin making preliminary plans for post-market monitoring and surveillance for both the AI-enabled medical device and its model(s) during this step. These plans can address the unique challenges of AI-enabled medical devices, such as performance degradation due to drift, and enable real-world performance evaluation. Regulatory requirements in the particular jurisdiction(s) in which the AI-enabled medical device will be deployed may help to scope what might be needed for this activity.

Additional Standards and References to be Considered

- ISO/IEC 23053:2022 *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

5.2. Data Collection and Management

Fit-for-purpose, representative[11] data is crucial for the appropriate design and development of AI models in medical devices, while also important for minimizing risks such as unwanted bias. While there is no shortage of recognized industry best practices for data collection and management, this step covers considerations for appropriate data collection and management practices necessary for training, tuning, and validation of AI models that can help support the model's safety and performance.

It is important for manufacturers to review the legal and regulatory requirements in the jurisdiction(s) in which data is created, collected, analyzed, secured, and stored (such as data privacy requirements and consent as well as requirements on data collection and data management) and follow good clinical practices[12], which are not extensively covered in this document. Furthermore, as it is common to outsource activities associated with data acquisition, tools, and services, it is important for manufacturers to consider recommendations in section *7.6 Managing Outsourced Processes, Activities and Products* in IMDRF/SaMD WG/N23 FINAL:2015 for these activities.

The rest of this section highlights key considerations relevant to data collection and management for AI models and AI-enabled medical devices. The recommendations included in this section can be applied across the AI-enabled medical device life cycle wherever data is being collected, used, and stored. This includes the processes that occur during the Verification and Validation life cycle step, for example, as well as for data collected in the post-market setting to maintain, monitor, or further train and improve model performance.

---

[10] Per IMDRF/AIMD WG/N67 (Edition 1):2022 *Machine Learning-enabled Medical Devices: Key Terms and Definitions*, "MLMD manufacturers, regulators, and users should be aware of the conflicting interpretations of the term validation and ensure that communication regarding the development phases and the associated datasets is clear to avoid confusion between data validation, ML model tuning, and medical device validation. Alternatively, the use of the term validation that refers to the training and tuning process should be avoided in the context of medical device development. It is recommended that the use of the term 'validation' be accompanied by the context when referring to ML model tuning, data curation, and the associated datasets".

[11] See GMLP Guiding Principle 3.

[12] See, for example, ISO 14155:2020 *Clinical investigation of medical devices for human subjects - Good clinical practice*

Data Suitability and Collection/Generation: While this may be generally addressed in the Planning and Design step, it is important for manufacturers to consider data suitability when further investigating sources of data that are readily available or need to be generated. It is important for data to align with the model's particular requirements, including input, intended patient population, disease state, and reference standard[13] and account for what may be needed to ensure adequate transparency to the end user (for more information on transparency, see Section 6 Transparency and Labelling). For example, for some use cases, age and sex data may not be needed to train the model but can help users understand performance across these demographics. When collecting and/or generating data, clearly defining inclusion and exclusion criteria is a critical activity to complete prior to collection or analysis (i.e., in the case of a retrospective data analysis). If multiple data sources will be used, for example a combination of a data repository and data from a prospective clinical study, it is important that there is consistency across these datasets (including, for example, data pre-processing), where possible, to ensure discrepancies will not cause issues in training or tuning.

Data Augmentation and Use of Synthetic/Simulated Data: As AI models become more complex and challenges arise with accessing data, manufacturers may use various methods, including data augmentation and synthetic/simulated data, to incorporate new and modified data into their datasets. It is important to note that the acceptability of these approaches depends on whether the jurisdiction in which the model and device are developed, validated, and deployed allows for such uses. Data augmentation techniques may be used in this content to modify existing datasets by applying transformations that preserve the underlying characteristics while increasing dataset size and diversity. For example, techniques such as applying controlled rotations, brightness adjustments, and geometric transformations to X-ray or MRI images can be used to increase dataset diversity while preserving clinically relevant features. Manufacturers may also seek to use synthetic or simulated data to supplement actual, human data for model development and validation. Use of synthetic or simulated data can help address ethical and privacy concerns and accelerate data generation, particularly for underrepresented populations such as patients with rare diseases.

Because data that is augmented and synthetic or simulated data may not accurately reflect the nuances and complexity of data that may be encountered in the real-world use of the model and device, using such data in AI model development and validation may introduce uncertainty, especially about whether synthetic data can serve as an adequate surrogate for human data in validation studies. In light of these concerns, it is important for manufacturers to carefully document how data was augmented or synthetically generated (considering both data provenance and transformation methods), how it is used, and document justification that the data and approach to its use is fit-for-purpose. Such documentation and justification can be particularly helpful to provide to regulatory authorities to help provide clarity on the appropriateness and risk-based approach taken to the use of such data. In addition, manufacturers can consider tagging such data as such at the point of generation to aid in traceability between, for example, synthetic and human-derived data sources.

Data Representativeness and Bias Mitigation: It is important that data collected and used for training, tuning and validation represents the full spectrum of data that the model is intended to encounter in clinical use in order to ensure that the data is fit-for-purpose. Furthermore, it is important that this data represents not only clinical subtypes, such as demographics of the intended use population and disease conditions (both cases and controls) but also nonclinical subtypes, such as data acquisition equipment and their parameters and collection sites relevant to the device's intended purpose.

[13] See GMLP Guiding Principle 5.

In addition to clinical and non-clinical subtypes, it is important for manufacturers to consider how under-represented and priority populations within the intended use population (e.g., specific ethnic groups, Indigenous populations, or rural and remote communities) are reflected in the data used for training, tuning, and validation. Generally, where adequate representation cannot reasonably be achieved, it is important to transparently document and relay these limitations to users, including potential impacts on performance and generalizability, and to implement appropriate mitigation measures such as additional local validation, cautious labelling, or risk-proportionate bridging evidence. Collecting data from multiple sources to represent the intended use population(s), environment(s), and context(s)-of-use may be a solution to potentially reduce under-representation as well as improve representativeness across other jurisdictions whilst reducing the need for subsequent bridging studies in this regard. However, it is important for manufacturers to evaluate different data sources to address specific potential risks of unwanted biases and confounding factors. For example, data from electronic health records (EHRs) recording routine health delivery may represent clinical decisions that take into consideration various factors related to a particular patient encounter such as cost and affordability issues or patient preferences. Hence, this data may not generalize to all health care settings.

Manufacturers can consider sampling strategies to ensure all relevant subgroups are appropriately represented in the dataset as well as to address potential algorithmic biases that may arise from use of specific datasets (e.g., sampling bias, selection bias, measurement bias, algorithm bias) through careful sampling, stratification, and bias mitigation techniques. When developing adaptive AI models that continue to learn after their initial deployment, it is important for manufacturers to eliminate or reduce as far as possible the risk of biased outputs influencing input for future operations (feedback loops), and to ensure that any such feedback loops are addressed with appropriate mitigation measures.

Data Cleaning/Quality Assurance: Data Cleaning, or the process of correcting and/or deleting incomplete, incorrect, or irrelevant records from a database or table,[14] may be needed to ensure data collected is fit-for-use. Trained professionals (e.g., domain experts, data scientists) and/or those with appropriate expertise are best suited to perform data cleaning to minimize introducing unwanted bias into the dataset. Furthermore, it is important that cleaned data be representative of the data the model will encounter in its intended use environment and, where applicable, that any anomalies or discrepancies identified during cleaning tasks inform risk control strategies as similar anomalies may be encountered with model inputs after deployment. Other key activities include prospectively defining data cleaning procedures with contextually relevant stop thresholds, setting limits on correcting errors (to be done to the extent possible without infringing on privacy), preventing duplication of data when aggregating from multiple sources, and maintaining documentation of all data cleaning and quality assurance steps to ensure reproducibility and traceability. Manufacturers may choose to implement automated processes for data extraction, transformation, and loading (ETL) to maintain data integrity while supporting scalable operations. This can include, for example, real-time data quality monitoring with automated anomaly detection to enable rapid identification of issues that could compromise model performance and patient safety. It is also important for manufacturers to validate data processing pipelines to ensure consistent and reproducible results.

Data Lineage and Provenance Documentation: Building upon the documentation recommendations mentioned for data cleaning, it is important for manufacturers to establish comprehensive data lineage and provenance tracking throughout the AI-enabled medical device life cycle to support traceability and remediation of data quality issues. Data lineage documents the data origin, transformations, and movement over time, providing visibility and providing manufacturers the ability to trace errors in device performance or other issues that occur in the post-market setting back to potential root causes. Data provenance documents the inputs, entities, systems, and processes involved in creating the data, providing a historical record of data origins.

[14] Artificial Intelligence-Data quality for analytics and machine learning (ML) – Part 4: Data quality process framework (ISO/IEC 5259-4)

Data Retention, Retirement and Life Cycle Management: Effective data life cycle management spans from initial collection through active use, archival storage, and final disposal, including well-defined criteria and processes for data retirement. Data use tracking mechanisms help monitor access patterns and identify potential overuse, which can warrant earlier retirement. Depending on the type of data, there may be distinct life cycle paths based on sensitivity, ownership, and intended use. While it is important for data to be retained in order to allow manufacturers to investigate adverse events, analyze device performance patterns, trace the source of malfunctions or safety concerns, and provide evidence for regulatory inquiries or corrective actions, retention periods typically reflect a balance of regulatory and legal requirements, clinical or research utility, and privacy obligations, as applicable to a particular jurisdiction. Robust governance frameworks that include documented policies, designated data stewards, and automated monitoring ensure consistent application of retention and disposal rules.

Additional Standards and References to be Considered

- ISO/IEC 5259-4:2024 *Artificial intelligence — Data quality for analytics and machine learning (ML) — Part 4: Data quality process framework*
- ISO/IEC 23053:2022 *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

5.3. Model Building and Tuning

In this next step of the life cycle, AI model(s) are developed and refined using the collected data, including selection of appropriate architecture, feature engineering, and performance optimization.

GMLP Guiding Principle 6 highlights that model choice and design are tailored to both the available data and the device's intended use, with design decisions evaluated to actively mitigate known risks like overfitting and performance degradation while supporting clinically meaningful performance goals. The rest of this section provides additional considerations relevant to model building and tuning for manufacturers to consider during this stage of development.

Model Design and Architecture Selection: Manufacturers are best positioned to make decisions when developing and tuning models, particularly for AI-enabled medical devices. As introduced in Section 5.1, design choices including feature selection, model type, and evaluation metrics, can inadvertently introduce or reinforce risks, such as inaccurate or biased outputs as described in Section 4.2. Avoiding overly complex models, e.g., limiting the number of parameters, layers, or features, may be optimal for addressing such risks, ensuring the model is fit-for-use and potentially reducing the amount of data needed during this step. Justification for model choices and rationale for decisions made with respect to the intended use of the device can not only support regulatory submissions but also help ensure that the device is and continues to be well-aligned in the clinical context and workflow across the life cycle. For instance, a developer working on a medical diagnostic AI-enabled medical device might explain their choice of a convolutional neural network (CNN) over a random forest model due to the CNN's superior performance in image recognition tasks, which is crucial for accurately identifying anomalies in medical scans.

Model Explainability and Interpretability: Building upon considerations for model selection described in Section 5.1, when choosing to employ a certain model, it is important for manufacturers to consider the explainability and interpretability of the model, including as it relates to the functionality of the AI-enabled medical device in which it is incorporated. Models that are explainable can enable users to understand their underlying reasoning and limitations and can also support the ability of the manufacturer to create an appropriate evaluation plan. However, model explainability may have potential trade-offs with complexity and performance. As such, manufacturers can consider the trade-offs between the potential risks and benefits of unexplainable models in the context of the clinical use, including the environment, the characteristics of the users, and the system into which the model will be deployed. There may also be a higher expected level of clinical evaluation for less explainable models to gain clinician trust for real-world adoption. By considering these topics during this step, manufacturers can avoid identifying challenges and risks later in development when it is difficult to address the underlying structure that might be leading to risks or reduced trust from the clinical community. A model's interpretability can also be considered to ensure that users will be able to comprehend how the model arrives at its outputs or recommendations.

Data and Feature Preprocessing: Data preprocessing is a critical step for optimizing data prior to model training. It includes activities such as data cleaning (described in Section 5.22 above), normalization, and transformation to ensure the data is consistent, reliable, and suitable for modelling. Regarding feature preprocessing, while not exhaustive, both feature engineering and selection as well as dimensionality reduction[15] are some of the many approaches that can be employed. When feature engineering and selection is used, it is beneficial for features within the raw data to be chosen with considerations for biological and scientific feasibility and evidence because these will inform clinical evaluation and explainability in terms that clinicians and patients can understand, providing a "valid clinical association"[16] as described in IMDRF/SaMD WG/ N41 Final:2017 *Software as a Medical Device (SaMD): Clinical Evaluation*. In the absence of full scientific evidence, it is possible that other steps of the life cycle may compensate for the lack of clarity in features selected, such as additional validation on a wide variety of datasets, which could be used to ensure the model performs as expected across the intended use population. Additionally, choosing features that are generalizable can help manufacturers avoid introducing unwanted bias for particular demographic groups and/or subgroups, including both clinical and non-clinical subtypes. For example, when choosing features for an AI-enabled medical device for predicting the risk of pre-term birth, manufacturers may want to select features that are biologically relevant and generalizable across populations, including factors such as previous pre-term births, maternal age, body mass index (BMI), and presence of certain infections. Finally, regardless of the approach to feature preprocessing used, it is important that manufacturers ensure the decision-making process of choosing the approach is transparent and justified, which can inform evidence for future regulatory inquiries and transparency for users.

Selection of Evaluation Metrics: It is important to select appropriate metrics to ensure reliable model performance. A loss function is the difference in error between the model's predictions and target (or reference) values and can be used during model training and building to guide adjustments to the model's parameters, such as weights and hyperparameters. An evaluation metric is used to evaluate a model's performance during tuning and can be used throughout the life cycle (when feasible) to ensure that the model is continuously meeting performance goals. Appropriate loss function and evaluation metrics chosen for model training, tuning and validation are relevant to the intended use and are generalizable to the target patient population. Selection of generalizable evaluation metrics may be particularly important when features could not be selected based on biological or scientific evidence. Examples of common evaluation metrics can be found in Appendix B.

[15] Dimensionality reduction helps to improve computational efficiency, reduce noise in the data, mitigate overfitting, and enhance interpretability of the model and is another option when performing data preprocessing.

[16] Per IMDRF/SaMD WG/ N41Final:2017 *Software as a Medical Device (SaMD): Clinical Evaluation*, "valid clinical association, also known as scientific validity, is used to refer to the extent to which the SaMD's output (concept, conclusion, measurements) is clinically accepted or well-founded (based on an established scientific framework or body of evidence), and corresponds accurately in the real world to the healthcare situation and condition identified in the SaMD definition statement.

Deployment Considerations: Specific requirements of the final deployment environment may result in considerations at the model tuning stage to address risks such as infrastructure inadequacy and computational and scalability constraints, as described in Section 4.2.4. For example, some clinical applications may pose hardware resource constraints or require fast inference times that limit the model design options. Such constraints may require, for example, compression/distillation of the initially designed model, creating smaller, more cost-efficient models by transferring knowledge from the initial, complex models. It is important that decisions regarding trade-offs inherent to addressing such design considerations be documented as part of the risk/benefit analysis.

Leveraging general-purpose and off-the-shelf models: If manufacturers choose to incorporate or leverage third-party, general-purpose models, including large language models (LLMs), other architectures or foundation models, into their AI-enabled medical devices, there are several factors to consider. For example, as with all medical devices, it is important to implement robust supplier management practices and follow industry best-practices for software of unknown provenance (SOUP)[17] and third-party oversight. Other key considerations include assessing the third-party supplier credibility, model provenance, model's longevity, and life cycle management processes, including the model's version history, update cadence, and support commitments. It is important for manufacturers to also assess and document the model's known limitations, potential biases, and uncertainties. When applicable, additional considerations include understanding the generative behaviour of the model, such as variability in outputs, susceptibility to hallucination, and dependence on input prompts or context.

When thinking about incorporating these types of models, it is important for manufacturers to evaluate feasibility of managing model and documentation updates, evaluate available information about the model while understanding how information gaps may impact safety, and implement appropriate risk controls, such as human oversight and guardrails, as well as those that may be needed for appropriate deployment of such models. Even when information is available on the training data set, manufacturers may find that general-purpose models are typically not trained on data that is representative of the intended use population or clinical use case. Manufacturers can consider using methods such as transfer learning and fine tuning as part of a comprehensive risk strategy to address these representativeness gaps. For generative AI-enabled models, it is also important to monitor for emergent behaviours or unintended content generation over time as the model or its environment changes.

Providing transparent information about the use of general-purpose and off-the-shelf models is important for both regulatory authorities and users to understand the foundation, reliability, and limitations of an AI-enabled device that incorporates these types of models. See Section 6 for more information on transparency and labelling for AI-enabled medical devices.

Additional Standards and References to be Considered

- Section 8.3 Development in IMDRF/SaMD WG/N23 FINAL: 2015
- ISO/IEC 23053:2022 *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*
- ISO/IEC TS 25058:2024 Systems and software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Guidance for quality evaluation of artificial intelligence (AI) systems
- ISO/IEC 25059:2023 Software engineering — Systems and software Quality Requirements and Evaluation (SQuaRE) — Quality model for AI systems

[17] IEC 62304:2015 defined the term as follows: "A software item that is already developed and generally available and that has not been developed for the purpose of being incorporated into the medical device (also known as "off- the-shelf software") or software previously developed for which adequate records of the development processes are not available. NOTE A MEDICAL DEVICE SOFTWARE SYSTEM in itself cannot be claimed to be SOUP"

- IEEE 2941-2021 IEEE Standard for Artificial Intelligence (AI) Model Representation, Compression, Distribution, and Management

5.4. Verification and Validation, including Clinical Evaluation

As with all medical devices, verification and validation (also referred to as "V&V") in the AI-enabled medical device life cycle builds trust in the performance of the device and establishes the baseline assurance that the device has met its design requirements and can be used safely and effectively for its intended purpose.

At this point in the life cycle, the model(s) have been built and tuned. Evaluation metrics (described in Section 5.3 with examples included in Appendix B) used as part of tuning can continue to establish the model and AI-enabled medical device's performance. It is important to note that same underlying model can be used to produce multiple types of outputs, which may be used, understood and therefore, evaluated in different ways. An AI model's outputs may include, for example, continuous numerical outputs (such as a thermometer that produces an estimate of core body temperature in degrees Celsius), categorical outputs (such as a thermometer that produces a result of "subnormal," "normal," or "fever") or binary outputs (such as a thermometer that produces a result of "positive" or "negative" for fever). The model's outputs will have a significant influence on the approach used in V&V processes to adequately evaluate the safety and effectiveness of an AI-enabled medical device. Furthermore, typically, models themselves are not deployed but are integrated into a device (i.e., an AI-enabled medical device) or system (e.g., hosted in cloud server) with additional functions that may impact the context or utility of the model outputs and will also impact what V&V strategies will be most appropriate.

This section will address considerations for V&V activities related to the model (Section 5.4.1), including external validation of the model, as well as considerations for V&V activities related to the AI-enabled medical device, including clinical evaluation[18] (Section 5.4.2). It is important to note that clinical evaluation distinguishes itself from external validation of the model, where clinical evaluation covers the entire medical device, including the intended user and workflow and not just the generalizability of the AI model(s).

It is also important for manufacturers to conduct verification and validation activities for an AI-enabled device in accordance with applicable design control processes for medical devices, (e.g., ISO 13485:2016). As part of these verification and validation activities, manufacturers may want to consider the recommendations in section 8.0 (SaMD Realization and Use Processes) of Software as a Medical Device (SaMD): Application of Quality Management System (IMDRF/SaMD WG/N23 FINAL:2015), as well as the definition of validation provided in Machine Learning-enabled Medical Devices: Key Terms and Definitions (IMDRF/AIMD WG/N67 (Edition 1):2022). Furthermore, IMDRF's GMLP principles provide some common principles that can be considered when validating medical devices that incorporate AI, such as using validation data that is independent from the data that was used to develop the model and is representative of the intended use population[19].

Documentation throughout the V&V processes is essential to track and monitor the model's strengths, limitations, and impact. This recordkeeping can help develop comprehensive and transparent documentation for the AI-enabled medical device and its model(s) so that clinicians and patients can understand the V&V outcomes and make informed choices regarding AI-enabled medical device deployment and use. See Section 6 for more information on transparency and labelling for AI-enabled medical devices.

[18] Per IMDRF/SaMD WG/N41FINAL:2017 *Software as a Medical Device (SaMD): Clinical Evaluation*, "clinical evaluation is a systematic and planned process to continuously generate, collect, analyze, and assess the clinical data pertaining to a SaMD in order to generate clinical evidence verifying the clinical association and the performance metrics of a SaMD when used as intended by the manufacturer" while "clinical validation measures the ability of a SaMD to yield a clinically meaningful output associated to the target use of SaMD output in the target health care situation or condition identified in the SaMD definition statement".

[19] See GMLP Guiding Principle's 3 and 4.

5.4.1. V&V Activities related to the Model

External Validation of the Model: Prior to widespread clinical evaluation and deployment, it is important to evaluate whether the model's performance generalizes beyond the training environment and remains performant in real-world clinical applications. This involves, for example, validating the model on a dataset that is separate from training data, and assessing the performance, accuracy, and reliability of the finalized AI model. It is also beneficial for external validation to address the socio-technical environment where the AI model will be deployed, as emphasized in the SaMD quality management framework, which requires "thorough understanding of the socio-technical environment (clinical perspective), and the technology and system environment (software perspective)" to prevent "incorrect, inaccurate, and/or delayed diagnoses and treatments" (IMDRF/SaMD WG/N23 FINAL:2015).

Model Robustness: Model robustness assesses the model's ability to continue to perform in conditions outside of the ideal use condition. Techniques such as stress testing and sensitivity analysis can be used to evaluate the model's resilience against variations in input data, noise, outliers, and adversarial attacks. Identifying such potential vulnerabilities early may allow for proactive adjustments that strengthen the model's stability and ensures that identified operational bounds and limitations can be conveyed to users to control for risks associated with improper or unsupported use. Additionally, red teaming, a process where a team intentionally probes the model for weaknesses and exploits potential vulnerabilities, is a valuable method for identifying issues and validating implemented risk controls and failure/error handling mechanisms. Demonstrating model robustness may allow for a better understanding of the risks related to device performance in edge cases and whether adequate risk controls are in place.

5.4.2. V&V Activities related to the AI-enabled Medical Device

Verification of the AI-enabled Medical Device: Once a manufacturer externally validates a model, it may be incorporated into an AI-enabled medical device and verified to ensure that requirements are met. In general, this can follow typical practices for medical device software. Additional verification activities may need to be performed according to medical device requirements.

Clinical Evaluation Study Design: It is important for manufacturers to ensure that the approach taken to clinically evaluate an AI-enabled medical device is tailored to its intended use and unique characteristics, including the intended population. As introduced above, clinical evaluation activities can depend on the outputs of the model and whether or not the model serves as the medical device's primary intended use or purpose or facilitates other device functionality. For example, in diagnostic medical devices like a radiology computer-aided detection or diagnosis medical device software, the model's output may be a prediction of whether and where evidence of cancer appears in an image. Diagnostic medical device evaluations may therefore focus specifically on determining whether the output of the model is accurate and clinically useful based on metrics like positive and negative percent agreement or predictive value. In contrast, for a therapeutic medical device such as an automated insulin delivery system, a model may use sensor and behaviour data (e.g., carbohydrates eaten, exercise planned) to estimate and predict a user's current and future blood glucose values, which may be used by another device function to determine how much insulin should be delivered to the user. Therefore, therapeutic device evaluations may focus more broadly on the therapeutic effectiveness of the overall device, including determining whether outcome-related measures were improved. Furthermore, the usability and evidence supporting the output, such as transparency on why a given output was provided, may also affect clinical outcomes and may be included in clinical evaluations where relevant. This is an especially important consideration for non-autonomous AI-enabled medical devices that function as decision support and rely on the user to make the final therapeutic, diagnostic or other clinical decision. When patient populations, disease characteristics and medical practices have similar documented characteristics across jurisdictions, it may in some cases be appropriate to extrapolate the clinical evaluation information to inform decisions in another appropriate patient population. High quality clinical evaluations that are consistent with the GMLP principles can reduce duplicative efforts, thereby further promoting and enabling regulatory reliance.

Finally, it is critical for the clinical evaluation study to consider the intended use and risk-level of the AI-enabled medical device with respect to the associated range of potential, appropriate strategies for validating its use, including from robust retrospective studies with suitable reference standards, silent mode studies, or simulation-based evaluations to prospective studies in the intended use environment. Where the device is intended to support or influence human clinical decision-making, it is important for the manufacturer to consider study designs that explicitly assess the performance and safety of the human-AI team[20].

Usability and Human Factors: To ensure that the AI-enabled medical device is effective and usable in its intended environment, it is important for manufacturers to consider what is needed for human factors and usability validation. This involves assessing whether the device aligns with the skills, expertise, and needs of its users, ensuring that they can correctly interpret and use its outputs and validate risk controls that are in place to address risks identified in Sections 4.2.1 and 4.2.2. Usability validation can involve structured strategies, such as silent testing, which evaluates how well the device integrates into real-world workflows while blinding clinicians to the outputs of the model so that the existing workflow is not disrupted. It is also essential to test the device's usability in the context of its intended purpose, validating that users can understand its inputs and outputs, recognize its limitations, and use it effectively for decision-making. This may include, for example, testing the ability of users to understand labelling, as further described in Section 6. It is also important for manufacturers to consider reasonably foreseeable misuse and whether any potential user errors could lead to unintended consequences.

Determining adequacy of measured performance: After conducting the V&V described above, the next step is to decide whether the measured performance is adequate. Judgements about whether or not medical device performance in V&V activities (or any other performance) is acceptable are influenced by several factors:

- the nature of the condition of interest, including whether the condition is relatively minor, is potentially deadly or debilitating, or is somewhere in between;
- the clinical workflow that the AI-enabled medical device is intended to be used within, particularly whether it functions within a system that includes prior screening or subsequent confirmations that can catch potential errors; and
- the usefulness, importance, and autonomy of the AI-enabled medical device outputs in the clinical environment, especially regarding whether a specific clinical action or intervention can be initiated based on the result and whether such interventions occur automatically without additional clinical oversight.

These judgements can become even more complex when an AI-enabled medical device is intended to be used repeatedly for the same individual over time. It is important to consider the limitations, uncertainty, and performance (including acceptance criteria) and whether they are acceptable for a given clinical use. Manufacturers may benefit from crafting a clear justification to both articulate to users on the benefits and risks of an AI-enabled medical device and as part of regulatory approval on why the V&V activities and outcomes are relevant to the device's intended use.

Additional Standards and References to be Considered

- ISO/IEC 23053:2022 *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

[20] See GMLP Guiding Principle 7.

5.5. Deployment

Deployment describes the process of integrating the AI-enabled medical device into its intended use environment, including "delivery, installation, setup, and configuration that support a controlled and effective distribution"[21] of the device, making it accessible for real-world use. This step demarks the transition into post-market and encompasses the processes between regulatory market authorization and first use of the device.

There are standards and resources that manufacturers can consider when planning for deployment requirements and processes, such as those described in ISO 13485, which emphasizes that deployment planning must be documented, risk-based, and include provisions for training, installation verification, and post-delivery activities to ensure safe and effective use of medical devices. It is important to note that the deployment process is highly context-specific to the device's intended use. For example, an AI-enabled medical device may be deployed in a consumer product for direct use by patients and caregivers or, conversely, it may be deployed into a complex hospital Information Technology (IT) network and interact with multiple stakeholders. Additionally, AI-enabled medical device performance can be significantly impacted by variations in IT infrastructure (e.g. scanner models and operating systems), patient demographics and local health policies, requiring deployment to be managed carefully and proportionate to the risks of the intended use.

It is important for manufacturers to review the relevant legal and regulatory requirements, industry standards, and organizational policies in the jurisdictions and sites in which the device is being deployed. Such considerations could include data privacy, security, and ethical considerations to maintain trust and mitigate risks in the deployment stage, which are not extensively covered in this document.

This section focuses on deployment considerations that are unique to or particularly important for AI-enabled medical devices.

Planning and Infrastructure: As part of the Planning and Design life cycle step as well as Model Building and Tuning, deployment is planned and infrastructure needed for deployment, operations and monitoring and real-world performance evaluation is designed (see Sections 5.1 and 5.3). The deployment step is meant to ensure that assumptions made about the device and its infrastructure during development were accurate in the real-world environment and provides an opportunity for adjustments to be made so that the device can be used safely and effectively. For instance, when a device is designed to integrate with an EHR system, deployment involves establishing and verifying this connection and, if data fields are labelled differently than anticipated during development, the system can be reconfigured to properly map inputs per the device's requirements. This can help to control for risks, for example those outlined in Section 4.2.4 related to interoperability issues.

As AI-enabled medical devices can have different failure modes than other medical devices and medical device software, it is important for this plan and associated infrastructure to include ways to determine whether any issues that arise can be attributed to AI model(s) or to medical device software generally, so that manufacturers can properly evaluate a failure and address it adequately. Additionally, it is important for the infrastructure to have mechanisms in place to determine whether the quality and quantity of inputs to and data requirements of the AI-enabled medical device meet specifications and account for variation across sites, enabling the ability to monitor for data drift and performance degradation, potentially through Human Oversight features (as described in Section 4.3). Furthermore, where applicable, it is important for the infrastructure to ensure that the clinician retains autonomy over the clinical decision-making including the ability to override AI-enabled medical device and its model(s)'s output where needed. If there is a plan to use real-world data to improve the model(s) in the future, it is important for that to be considered when building and implementing the infrastructure and principles and practices from the Data Collection and Management life cycle step can be employed. Finally, it is essential for the deployment plan and infrastructure to account for cybersecurity concerns detailed in Section 4.4.

[21] IMDRF/SaMD WG/N23 FINAL:2015

Deployment Verification and Validation: Deployment verification and validation ensures that the deployment plan was executed correctly and any anomalies that may have occurred during deployment are appropriately addressed before widespread use of the AI-enabled medical device. This can include ensuring that training for site staff and healthcare professionals was adequate and appropriate for the environment and users. Depending on the risk of the device, deployment V&V activities can range in complexity and timespan. For example, utilizing a combination of site-specific retrospective data and control data may be sufficient to verify that the deployment process was successful. However, complex workflows, which may contain highly variable data inputs, can require more longitudinal activities such as trial periods with sufficient human oversight to validate the intended purpose of the AI-enabled medical device is being achieved.

Deployment Approach: Defining the implementation approach in deployment plans is a critical step, whether that is deployment in phases, such as a preview release followed by a final release, or through phased release across multiple sites. A phased release across multiple sites may be beneficial for AI-enabled medical devices because it enables manufacturers to identify potential safety issues, performance variations, or integration challenges in a controlled manner before full-scale deployment, allowing them to understand the AI-enabled medical device's performance across patient populations, clinical workflows, and technical environments that may differ from the controlled conditions used during V&V.

Site-Specific Customization and Localization[22]*:* AI-enabled medical devices may have the ability to be customized and/or localized to achieve expected performance levels at a particular site. Depending on the risks and context of the deployment environment, various approaches may be utilized. For example, parameters may need to be adjusted from a range of previously validated options until the most appropriate configuration for a deployment site is found (customization). Alternatively, or potentially in addition, local data may be used to retrain the AI model contained within an AI-enabled medical device to more appropriately align to site-specific criteria (localization). For devices that are localized, deployment includes localization steps and the development of training and labelling materials specific to each deployment. Predetermined Change Control Plans (PCCPs) can be leveraged and applied to AI-enabled medical devices as a way for regulatory authorities within a specific jurisdiction to prospectively review and approve certain changes to an AI-enabled medical device, including customization and localization, where such an approach is available.

Traceability and Version Control: For AI-enabled medical devices, especially those that are customized or localized as described above, it is important for traceability and version control mechanisms to be in place so that manufacturers have the ability to identify failure events and root causes and ultimately intervene as necessary to minimize risk to patients, such as those identified in Section 4.2.4 Risks Related to Deployment and Post-Marketing Monitoring and Performance. Such interventions may include retraining, which can be enabled by a PCCP, or rolling back to prior versions. Unique device identifiers (UDI), as recommended in IMDRF/UDI (Unique Device Identifier) WG/N48 FINAL:2019 *Unique Device Identification system (UDI system) Application Guide,* can be employed for software/model version traceability, where available. If UDIs are not applicable in a particular jurisdiction, manufacturers may want to consider alternative auditable identifiers.

Additional Standards and References to be Considered

- IMDRF/SaMD WG/N23 FINAL:2015 *Software as a Medical Device (SaMD): Application of Quality Management System*
- ISO/IEC 23053:2022 *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

[22] Please note that related changes to MLMD are described in Section 7.1.1 of IMDRF/AIMD WG/N67 (Edition 1):2022 Machine Learning-enabled Medical Devices: Key Terms and Definitions as "[h]eterogeneous changes [that] are non-uniform changes that can be specific to one clinic, region, demographic, etc. (sometimes referred to as local adaptations)". Both terms used in this section (localization and customization) are examples of these types of changes and the methods used to facilitate such changes.

5.6. Operations and Monitoring

As with all medical devices, monitoring of deployed AI-enabled medical devices can help detect and address any performance[23], technical, security and operational issues arising in the intended use environment.

IMDRF/SaMD WG/N23 FINAL:2015 recommends that “post market surveillance including monitoring, measurement and analysis of quality data can include logging and tracking of complaints, clearing technical issues, determining problem causes and actions to address, identify, collect, analyse, and report on critical quality characteristics of products developed". ISO 13485:2016 includes similar provisions for post-market surveillance and feedback.

The legal responsibility to meet medical device requirements, including post-market monitoring and surveillance, falls on the manufacturer of the AI-enabled medical device; in practice, manufacturers may depend upon the deploying entity/site to carry out post-market responsibilities, including collection of much of the data needed to track the performance of their device. Throughout this step of the life cycle (and Real-World Performance Evaluation in Section 5.7), it is important for manufacturers to consider how responsibilities may need to be shared and communicated to the user.

The rest of this section covers aspects of operations and monitoring that are specific or unique to AI-enabled medical devices, many of which enable the real-world performance evaluation described in the section below. Many, if not all, of the systems that facilitate operations and monitoring help to facilitate real-world performance evaluation. In this document, “Operations and Monitoring” refers primarily to continuous, often automated activities embedded within the quality management system (for example, logging, alerting, and incident handling), whereas “Real-World Performance Evaluation” (Section 5.8) refers to structured analyses or studies that use these data, and where appropriate additional data collection, to answer specific questions about safety, effectiveness, and clinical utility. Although concepts may overlap between these two sections, this distinction can help clarify roles and expectations across the AI-enabled medical device life cycle.

Infrastructure, Data and Logging: As discussed in Section 5.5 above, the deployed AI-enabled medical device’s infrastructure facilitates the collection of data needed to fulfill monitoring requirements, and can incorporate data logging and performance tracking capabilities, as well as anomaly detection algorithms, incident response procedures and troubleshooting workflows. For AI-enabled medical devices, it may be worthwhile for the infrastructure to have both local monitoring capabilities as well as the ability to contribute accumulated monitoring data to central repositories, ensuring a comprehensive and uninterrupted record of the AI-enabled medical device’s operational performance across deployments for the manufacturer. Where possible, the process may integrate advanced technologies that enable continual assessment of the AI-enabled medical device’s accuracy and responsiveness, supporting rapid intervention, adaptive recalibration, and maintenance of the AI-enabled medical device’s reliability and effectiveness across its operational life cycle. In addition, logging and auditing are critical components of an AI-enabled medical device’s operational monitoring activities and maintenance. For AI-enabled devices specifically, it may be beneficial for the logging mechanisms to capture detailed logs and audit trails (as allowed within jurisdictional regulations) of model predictions and any explainability elements (if available) such as heatmaps or top predictive features, data inputs, model outputs, and metadata such as timestamps, user interactions, model versions, and environmental context for traceability, accountability, and compliance purposes.

[23] See GMLP Guiding Principle 10.

Performance Degradation and Drift Detection: Performance of the models within the AI-enabled medical devices can degrade over time as the statistical patterns that informed the original model(s') training may become less representative at one or more deployment sites, causing the AI-enabled medical devices to generate increasingly inaccurate or inappropriate outputs, even though the model(s) have not changed. Drift can emerge from changes in input data characteristics, shifts in underlying population dynamics, changes in context of use, such as clinical behaviour or patient management, or gradual accumulation of data biases that were not apparent during initial training. Continuous post-market monitoring and surveillance enables early detection of these performance variations, yielding better visibility to manufacturer and supporting the implementation of timely recalibration, retraining, or intervention strategies when necessary and with appropriate regulatory approval. If drift is detected, it may be important to consider additional investigations, such as retrospective analysis. Monitoring drift in generative AI-enabled medical devices is especially critical as these systems can rapidly and unpredictably deviate from their original training objectives. Effective drift monitoring ensures that generative AI-enabled devices maintain their core functional integrity and alignment with the original design. The non-deterministic nature of some AI models creates unique challenges for post-market performance assessment, as outputs may vary from those observed during pre-market testing without necessarily indicating improved or degraded performance. In these cases, manufacturers can consider developing post-market metrics that account for output variability and assess whether the model's performance remains within acceptable bounds and implement monitoring strategies that distinguish between expected variation and performance drift.

Alerting and Reporting: The use of continuous monitoring and responsive alerting and reporting systems helps to ensure reliability, stability and effectiveness of AI-enabled medical devices at deployed sites. For example, use of thresholds for critical performance metrics, automated notifications and alerts are some of the ways manufacturers can continuously monitor their deployed AI-enabled medical devices and promptly address risks. In addition to individual incidents, it is important to monitor for changes in data distribution and performance over time, e.g., drift as described above. Jurisdictional requirements may dictate relevant alerting and reporting needs to ensure timelines and other regulatory requirements for reporting are met.

Monitoring for Advanced AI Models: Operations monitoring for customizable and adaptive AI-enabled devices may necessitate fundamentally different approaches due to their inherent architectural differences. Monitoring customizable AI-enabled medical devices, which operate within predefined parameters defined as part of regulatory authorization and set during deployment, focuses on verifying consistent performance within established boundaries and detecting any deviations from original specifications. Adaptable AI-enabled medical devices, which can have their algorithms and learning processes modified, necessitate more dynamic and sophisticated monitoring strategies that track not only performance metrics but also the AI-enabled medical device’s own learning and transformation mechanisms. This may include continuous surveillance of the adaptable device’s decision-making evolution, potential drift in core algorithmic logic, and ongoing assessment of whether the AI-enabled medical device’s modifications maintain alignment with the approved design. The monitoring approach for adaptable AI-enabled medical devices is inherently more complex, incorporating real-time analysis of the AI-enabled medical device’s learning trajectories, unexpected behaviour emergence, and potential unintended consequences of autonomous algorithmic modifications.

In addition, operational monitoring of AI-enabled medical devices such as generative AI-enabled medical devices may necessitate different post-market monitoring due to their emergent behaviour, user-driven variability, and context-sensitive outputs. To effectively capture unexpected uses, dynamic, usage-aware, and risk-responsive monitoring approaches may be beneficial, including approaches that monitor how the AI-enabled medical device is used due to the possibility of the AI-enabled medical device being repurposed, misused, or used beyond its intended use. Tracking the deployment context and collecting interaction logs and prompt-response histories from LLM’s (with privacy safeguards) as well as employing mechanisms to detect patterns of prompt chaining and workarounds may be helpful to flag unsafe deviations. Prompt classification and risk-based escalation to overseeing AI-enabled medical devices or manual review may be helpful mechanisms to implement in order to monitor and address developing unsafe uses.

Additional Standards and References to be Considered

- ISO/IEC 23053:2022 *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

5.7. Real-World Performance Evaluation

Understanding how an AI-enabled medical device performs in the real world[24] can be beneficial to supporting clinical utility and better understanding of the risks and benefits of the AI-enabled medical device. It can also facilitate the collection of data at scale as compared to pre-market evaluation, and help manufacturers understand how the device is performing in subpopulations in detail. It is important for manufacturers to consider implementing a structured and strategic approach to establishing and tracking the real-world performance of AI-enabled medical devices.

Monitoring infrastructure and data described in the sections above enables real-world performance evaluation described in this section (e.g., detailed logs and audit trails of model predictions and explainability elements). Drift monitoring in this step goes beyond the individual model's performance to understand how the AI-enabled medical device and its model(s) are performing in the context of the clinical workflow. For example, if clinicians are frequently choosing to override aspects of the AI-enabled medical device's outputs or the output entirely (although the Operations and Monitoring infrastructure indicates that the model is performing as expected), it may be worth the manufacturer looking into this further as part of the performance evaluation in its clinical context.

The rest of this section provides an overview of considerations with respect to real-world performance evaluation of AI-enabled medical devices.

Define Device Performance Indicators: It is important for manufacturers to establish device performance indicators that are framed by the intended use of the device/model and that are both informative and scalable. Ideal device performance indicators reflect the safety and effectiveness of the AI-enabled medical device in its real-world setting and allow for the detection of post-market safety signals. Depending on the AI-enabled medical device and model, manufacturers may wish to consider multiple different device performance indicators to capture different failure modes and different aspects of safety and performance, some of which could be the evaluation metrics used in prior steps of the life cycle when feasible. For example, device performance indicators can include technical measures of performance such as accuracy, precision, recall, and latency or other measures that capture the performance of the human-AI team such as user satisfaction.

Where feasible, it is important for device performance indicators to be defined not only in aggregate but also for clinically relevant subgroups, such as demographics, disease subtypes, and deployment sites, to support the identification of differential performance that could impact safety, effectiveness, or equity. It is also helpful to pre-specify risk-based thresholds, trends, or trigger conditions for these indicators (for example, sustained degradation in a safety-critical metric or emergence of significant performance gaps between subgroups) that may prompt investigation, mitigation, or model updates, in line with applicable regulatory requirements and change control mechanisms.

For AI-enabled medical devices enabled by adaptable and/or autonomous models, it is important for manufacturers to specify device performance indicators that measure whether the model operation remains within an acceptable performance range. Manufacturers may also consider defining governance procedures for remediation when model performance falls outside of the specified acceptable range according to predetermined criteria (e.g., sustained performance degradation exhibited with a particular frequency over a specified time span).

Data Collection: The infrastructure described in Section 5.5. and 5.6 can enable collection of necessary and representative data to measure the device performance indicators, with efforts made to ensure that data is of sufficient quality and integrity and is anonymized and protected, where applicable as required by jurisdictional regulations. Additionally, proactive data collection can occur outside of this infrastructure as well, such as interviews with clinicians and other users and surveys to collate actionable information.

[24] See GMLP Guiding Principle 10.

Optimizing Performance: While performance optimization can occur based on signals that surface during Operations and Monitoring, insights from real-world performance evaluation can also drive continuous improvement in the AI-enabled medical device, including when device performance indicators demonstrate consistent decline, when safety or effectiveness of the AI-enabled medical device is compromised, when significant technological or methodological improvements emerge, or from user feedback. This may lead to, for example, performance optimization by retraining the model(s), and considerations throughout this document can inform that activity, along with an assessment of the need for new regulatory approval.

Additional Standards and References to be Considered

- ISO/IEC 23053:2022 Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)

5.8. Sunsetting

Sometimes called decommissioning, retirement, or end-of-life, as discussed in IMDRF/SaMD WG/N23 FINAL:2015 for a SaMD, sunsetting involves the termination of "maintenance, support, and distribution" of the AI-enabled medical device "in a controlled and a managed fashion." In common with SaMD, "this process indicates an end to active support, and may entail deactivation and/or removal of [the AI-enabled medical device] and its supporting data." This can include, for example, technical decommissioning, which can involve gradually scaling down and eventually deactivating system components, ensuring a controlled shutdown to prevent data loss or service disruptions. It can also include decommissioning and repurposing of hardware and cloud resources used by the AI-enabled medical device. For AI-enabled medical devices specifically, it is important to communicate sunsetting plans to stakeholders, both internal teams and external users and sites, with clear timelines and reasons for the decision, as well as alternative solutions or support to help them transition, as applicable. Furthermore, as described in the Data Collection and Management life cycle step, retention of data for regulatory, legal and business reasons is critical, ensuring compliance with jurisdictional data protection regulations. It is important for manufacturers to maintain detailed documentation of the sunsetting process for audit purposes and to address any future legal or compliance inquiries.

Additional Standards and References to be Considered

- ISO/IEC 23053:2022 *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

6. Transparency and Labelling

Transparency, or the communication of clear, essential information[25] about an AI-enabled medical device to relevant audiences, is a key principle throughout the life cycle of an AI-enabled medical device and can help address risks related to information (see Section 4.2.1) and risks related to the human-AI interaction (see Section 4.2.2).

IMDRF/GRRP WG/N52: FINAL:2024 (Edition 2) *Principles of Labelling for Medical Devices and IVD Medical Devices* provides "general labelling principles, including specific sections on the label, instructions for use, and information intended for the patient" for all medical devices. These apply when developing labelling for AI-enabled medical devices as well, in addition to jurisdictional requirements for medical device labelling. As with all medical devices, AI-enabled medical devices should be accompanied by information that is truthful, accurate, and sufficient to support their use by the intended user and to ensure that they are safe and effective for their intended use or purpose. This information helps ensure, for example, users have the information they need about both the characteristics of the AI-enabled medical device, such as its model(s), and how those characteristics were assessed. See Appendix C for the types of information to be included in labelling to support safe and effective use.

With AI-enabled medical devices, it is particularly important that necessary information is presented to the user at the most appropriate time and in the most appropriate location and manner during medical device procurement, implementation, use, modifications, or at the AI-enabled medical device's end-of-life. When appropriate, labelling may sometimes be implemented through the medical devices' user interface (UI), with additional information included in the UI that can help further achieve transparency. This can include information related to the explainability and interpretability of the AI-enabled medical device's model behaviour and its output(s). For example, it is important for users who will rely on the output of the AI-enabled medical device in time-critical decision making to have a clear understanding of the relationship between how the safety and effectiveness of the device was demonstrated with respect to the type of patient and condition being treated (including regulatory authorization information as applicable) and when it is appropriate to rely on the output in these situations. It may impede clinical decision-making, if the user needs to interpret complex model outputs without accompanying clarifying information, especially in emergency situations, and overburden these users. Manufacturers may also wish to consider electronic labelling, which can support increased availability, utility, interactivity, and accessibility of labelling.

For AI-enabled medical device models with no planned updates, along with the types of information recommended within Appendix C, it is important that transparency efforts also focus on the model's initial technological and performance characteristics. For models that are expected to undergo discrete updates, tailored transparency information can be provided at the time of each update, including whether the update required regulatory approval, for example as part of PCCP. For autonomously adaptable models designed to perform self-implementing updates, transparency may also include details on the update process and the parameters or objectives guiding those updates. When general-purpose, off-the-shelf models are incorporated (as described in Section 5.3), it is important for users to have information on these models and their limitations. In all cases, it is important for the manufacturer to ensure the intended user can understand what, if any, risk controls and monitoring are in place and what, if anything, is expected of them to ensure safe and effective use on an ongoing basis.

Additional Standards and References to be Considered

- Transparency for Machine Learning-Enabled Medical Devices: Guiding Principles (Published June 2024): https://www.fda.gov/medical-devices/software-medical-device-samd/transparency-machine-learning-enabled-medical-devices-guiding-principles

[25] See GMLP Guiding Principle 9.

7. Conclusion

In summary, this document provides a harmonized framework to address the challenges and opportunities of AI-enabled medical devices, ensuring safety, effectiveness, and patient-focused innovation. The universal concepts of quality and risk management, cybersecurity, human oversight (see Section 4) are essential for their responsible development and use. The Life Cycle Steps (see Section 5) help manufacturers to ensure that AI-enabled medical devices and their model(s) are designed, developed, and deployed with GMLP, patient safety and clinical effectiveness in mind.

This document complements existing IMDRF publications and regulatory requirements, offering a common language and principles adaptable to local contexts. Jurisdictions may integrate these concepts to promote global consistency and alignment.

The goal is to foster innovation while ensuring AI-enabled medical devices are safe, effective, and capable of delivering meaningful benefits to patients and healthcare systems. By adopting the considerations and principles outlined in this document, stakeholders can contribute to the development of AI-enabled medical devices that are trustworthy, transparent, and aligned with the needs of patients, healthcare providers, and requirements of regulatory authorities. This collaborative approach will help ensure that AI-enabled medical devices continue to advance healthcare while maintaining the highest standards of safety and effectiveness.

Appendix A: Traceability between GMLP and Life Cycle Steps/Document Sections

AI Life Cycle Step and/or Document Section	Relevant GMLP Principle
Planning and Design	All
Data Collection and Management	Guiding Principle 3: Clinical evaluation includes the use of datasets that are representative of the intended patient population
	Guiding Principle 4: Training datasets are independent of test sets
	Guiding Principle 5: Selected reference standards are fit-for-purpose
Model Building and Tuning	Guiding Principle 2: Good software engineering, medical device design, and security practices are implemented throughout the total product life cycle
	Guiding Principle 6: Model choice and design are tailored to the available data and the intended use/ intended purpose of the device
Verification and Validation, including Clinical Evaluation	Guiding Principle 3: Clinical evaluation includes the use of datasets that are representative of the intended patient population
	Guiding Principle 4: Training datasets are independent of test sets
	Guiding Principle 5: Selected reference standards are fit-for-purpose
	Guiding Principle 7: The device is assessed with a focus on human-AI interactions in the intended use environment, including the performance of the human-AI team, rather than just the device in isolation
	Guiding Principle 8: Testing demonstrates device performance during clinically relevant conditions
Deployment	Guiding Principle 10: Deployed models are monitored for performance and re-training risks are managed
Operations and Monitoring	Guiding Principle 10: Deployed models are monitored for performance and re-training risks are managed
Real-World Performance Evaluation	Guiding Principle 10: Deployed models are monitored for performance and re-training risks are managed
Sunsetting	Guiding Principle 9: Users are provided clear, essential information

Transparency and Labelling	Guiding Principle 9: Users are provided clear, essential information

1179

IMDRF International Medical Device Regulators Forum

Appendix B: Examples of Common Evaluation Metrics

Examples of common evaluation metrics may include, but are not limited to, the following:

- Accuracy (measures the overall correctness of the predictions);
- Precision (measures the proportion of true positive predictions out of all positive predictions);
- Recall (measures the ability to correctly identify actual positives);
- Model performance, including different classification thresholds and discrimination ability;
- F1 score (measures precision and recall);
- Receiver Operating Characteristic – Area Under Curve (ROC-AUC) (evaluates model performance different classification thresholds and discrimination ability); and
- Mean squared error (MSE) (measures differences between predicted and actual values).

Other metrics like perplexity (metric to measure the uncertainty of a language model when predicting the next word in the sequence) and coherence (assesses the contextual relevance and logical structure of generated text) may be used for evaluation of LLM-based models.

Appendix C: Labelling Elements

For AI-enabled medical devices, information that may be needed to support safe and effective use includes, but is not limited to:

- the intended use or purpose and intended user;
- any professional or other qualifications that are necessary to use the device safely and effectively;
- clear information on compatibility requirements for inputs (if applicable);
- how the device should be used, including any prerequisite steps or necessary preparations;
- the populations in which the device can be used and the conditions for which it can be used (including known limitations);
- measured performance and subgroup performance;
- when and for how long the device should be used;
- pre-market testing conditions and datasets;
- how to interpret the output of the device, known limitations, model description or characteristics.

See also IMDRF/GRRP WG/N52 FINAL:2024 (Edition 2) *Principles of Labeling for Medical Devices and IVD Medical Devices* for additional information to consider.

1223

Please visit our website for more details.

www.imdrf.org