IMDRF International Medical Device Regulators Forum

Final Document

IMDRF/AIML WG/N88 FINAL: 2025

Good machine learning practice for medical device development: Guiding principles

AUTHORING GROUP

Artificial Intelligence/Machine Learning-enabled Working Group

27 January 2025

Preface

© Copyright 2025 by the International Medical Device Regulators Forum.

This work is copyright. Subject to these Terms and Conditions, you may download, display, print, translate, modify and reproduce the whole or part of this work for your own personal use, for research, for educational purposes or, if you are part of an organisation, for internal use within your organisation, but only if you or your organisation do not use the reproduction for any commercial purpose and retain all disclaimer notices as part of that reproduction. If you use any part of this work, you must include the following acknowledgement (delete inapplicable):

All other rights are reserved and you are not allowed to reproduce the whole or any part of this work in any way (electronic or otherwise) without first being given specific written permission from IMDRF to do so. Requests and inquiries concerning reproduction and rights are to be sent to the IMDRF Secretariat.

Incorporation of this document, in part or in whole, into another document, or its translation into languages other than English, does not convey or represent an endorsement of any kind by the IMDRF.

Naoyuki YASUDA, IMDRF Chair

Contents

Introduction

Artificial intelligence (AI) technologies, including machine learning, have the potential to transform health care by deriving new and important insights from the vast amount of data generated in health care every day. They use algorithms that can learn from real-world use and potentially use this information to improve the product's performance. But they also present unique considerations due to the iterative and data-driven nature of their development. This document establishes a common set of principles for the community to promote the development of safe, effective, and high-quality medical devices that incorporate AI. These principles are important to apply across the lifecycle of the medical device.

The 10 guiding principles for Good Machine Learning Practice (GMLP) presented in this document are a call to action to international standards organizations, international regulators, and other collaborative bodies to further advance GMLP. Areas of collaboration include research, creating educational tools and resources, international harmonization, and consensus standards, to inform regulatory policies and regulatory guidelines. These guiding principles may be used to adopt practices from other sectors, tailor them to the medical technology and healthcare, and to develop novel practices for this domain.

Further advances in AI technologies in healthcare, exemplified by generative AI, highlight the importance of clearly describing a product's intended use/ intended purpose and identifying its regulatory status. Moreover, generative AI may heighten the importance of GMLP, including fundamental software engineering practices. For example, healthcare technologies that incorporate generative AI may employ foundation models that are not under the provenance of the medical device manufacturers, thereby potentially introducing unique risks. Generative AI may also pose a more fundamental challenge with respect to demonstrating device performance. The regulatory science of measuring performance as well as characterizing and detecting errors in these models is maturing to meet this challenge.

As the AI medical device field continues to evolve, so too must GMLP and consensus standards. Strong partnerships with our international public health partners are essential to empower responsible innovations in this area. Thus, we expect this collaborative work can inform future IMDRF efforts and other international engagements.

References

IMDRF/SaMD WG/N10 FINAL:2013 *Software as a Medical Device (SaMD): Key Definitions*

IMDRF/SaMD WG/N12 FINAL:2014 *Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations*

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

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

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

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

IMDRF/CYBER WG/N70 FINAL:2023 (Edition1) *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/MC/N79 DRAFT: 2023 *Guiding Principles to Support Medical Device Health Equity*

IMDRF/SaMD WG/N81 DRAFT:2024 *Medical Device Software: Considerations for Device and Risk Characterization*

Guiding principles

1. **The intended use/ intended purpose of the device is well understood, and multi-disciplinary expertise is leveraged throughout the total product life cycle**: In-depth understanding of a medical device's intended use/ intended purpose[1] including context of use within the clinical workflow, and the desired benefits and associated patient risks, can help ensure that AI-enabled medical devices[2,3] address clinically meaningful needs over the total product life cycle of the device[4]. Multi-disciplinary expertise provides context-specific insight and experience, informs the intended use/ intended purpose, and enhances the safety and effectiveness of the device.

2. **Good software engineering, medical device design, and security practices are implemented throughout the total product life cycle**: Model design is implemented and maintained with attention to the fundamentals: robust software engineering practices, usability, data quality assurance, data management, cybersecurity[5,6,7], and quality management practices[4]. These practices include methodical risk management[8] and design processes that can appropriately record and communicate decisions and rationale, as well as ensure traceability, reproducibility, data authenticity, confidentiality, integrity, and availability. The infrastructure needed for model deployment, monitoring, and maintenance is carefully considered. These practices help support the rights, safety, and welfare of patients, including through the ethical use of patient data.

---

[1] IMDRF/SaMD WG/N81 DRAFT:2024 *Medical Device Software: Considerations for Device and Risk Characterization*

[2] IMDRF/SaMD WG/N10 FINAL:2013 *Software as a Medical Device (SaMD): Key Definitions*

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

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

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

[6] IMDRF/CYBER WG/N70 FINAL:2023 (Edition1) *Principles and Practices for the Cybersecurity of Legacy Medical Devices*

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

[8] IMDRF/SaMD WG/N12 FINAL:2014 *Software as a Medical Device: Possible Framework for Risk Categorization and Corresponding Considerations*

3. **Clinical evaluation includes the use of datasets that are representative of the intended patient population**: Data collection protocols aim to ensure that the relevant characteristics of the intended patient population (for example, in terms of age, gender, sex, race, ethnicity, geographical location, medical condition)[9], intended use environment, and measurement inputs are sufficiently represented in a sample of adequate size in the datasets for training, testing, and monitoring so that results can be reasonably generalized to the intended population of interest. These are fundamental for clinical evaluations[10] and important to manage any unintended bias[3] or dataset drift, promote appropriate and generalizable performance across the intended patient population, assess usability, and identify circumstances and subgroups where the model may underperform including over time.

4. **Training datasets are independent of test sets**: Training and test datasets[3] are selected and maintained to be appropriately independent of one another. All potential sources of dependence, including factors related to patients, sites, and data acquisition, are considered and addressed to assure independence. The extent of external validation is proportionate to risk.

5. **Selected reference standards are fit-for-purpose**: Accepted methods for developing fit-for-purpose reference standard[3] ensure that clinically relevant and well characterized data are collected and that the limitations of reference standards are understood. This includes documentation of the rationale for the choice of reference standards based on the device's intended use/ intended purpose and assessment of their suitability to address the intended use environment. If available, accepted reference standards in model development and testing that promote and demonstrate model robustness and generalizability across the intended patient population are used. The selection of reference standards is based on broad consensus, where available, and appropriate expertise.

6. **Model choice and design are tailored to the available data and the intended use/ intended purpose of the device**: Model choice and design are evaluated and shown to be suited to the available data and support the active mitigation of known risks, like overfitting, performance degradation, and security risks. The clinical benefits and risks related to the product are well understood, used to derive clinically meaningful performance goals for testing, and support the product's safety and effectiveness in achieving its intended use/ intended purpose[1]. Considerations include the impact on both the overall intended patient population and its subgroups as well as uncertainty and variability in the device inputs, outputs, and clinical use conditions.

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.** The performance of the device is assessed in the context of the intended use environment and clinical workflow, considering interactions with health care providers, patients, and caregivers where applicable. Human factors considerations are addressed, including for example, user skills, user expertise, user understanding of the model outputs and limitations,

[9] IMDRF/MC/N79 DRAFT: 2023 *Guiding Principles to Support Medical Device Health Equity*

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

potential for overreliance, level of device autonomy, and user error, for normal use and reasonably foreseeable misuse.

8. **Testing demonstrates device performance during clinically relevant conditions**: Methodologically and statistically sound test plans are developed and executed to generate clinically relevant device performance[10] information independently of the training dataset. Considerations include the intended patient population, relevant subgroups, clinical environment and use by the human-AI team, measurement inputs, and potential confounding factors.

9. **Users are provided clear, essential information**: The intended audience (such as health care professionals or patients) are provided clear, contextually relevant information appropriate to their needs. This includes the product's intended use/ intended purpose[1] and indications for use, benefits and risks, performance of the model for appropriate subgroups, the study methodology, characteristics of the data used to train and test the model, acceptable inputs, known limitations, user interface interpretation, clinical workflow integration of the model, and to the extent possible the basis for model output. Users are also made aware of the scope and timing of device modifications and updates. They are provided a means to communicate product concerns to the manufacturer.

10. **Deployed models are monitored for performance and re-training risks are managed**: Deployed models have the capability for an appropriate level of ongoing monitoring in "real world" use with a risk-based focus on maintained or improved safety and performance[4,10]. Additionally, when models are retrained after deployment, there are appropriate controls in place to manage risks of overfitting, unintended bias, or degradation of the model (for example, dataset drift) that may impact the safety and performance of the model.

Please visit our website for more details.

www.imdrf.org

Disclaimer

© Copyright 205 by the International Medical Device Regulators Forum.

This work is copyright. Subject to these Terms and Conditions, you may download, display, print, translate, modify and reproduce the whole or part of this work for your own personal use, for research, for educational purposes or, if you are part of an organisation, for internal use within your organisation, but only if you or your organisation do not use the reproduction for any commercial purpose and retain all disclaimer notices as part of that reproduction. If you use any part of this work, you must include the following acknowledgement (delete inapplicable):

All other rights are reserved, and you are not allowed to reproduce the whole or any part of this work in any way (electronic or otherwise) without first being given specific written permission from IMDRF to do so. Requests and inquiries concerning reproduction and rights are to be sent to the IMDRF Secretariat.

Incorporation of this document, in part or in whole, into another document, or its translation into languages other than English, does not convey or represent an endorsement of any kind by the IMDRF.