Good Machine Learning Practice for Medical Device Development: Guiding Principles

October 2021

The U.S. Food and Drug Administration (FDA), Health Canada, and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) have jointly identified 10 guiding principles that can inform the development of Good Machine Learning Practice (GMLP). These guiding principles will help promote safe, effective, and high-quality medical devices that use artificial intelligence and machine learning (AI/ML).

Artificial intelligence and machine learning technologies have the potential to transform health care by deriving new and important insights from the vast amount of data generated during the delivery of health care every day. They use software algorithms to learn from real-world use and in some situations may use this information to improve the product's performance. But they also present unique considerations due to their complexity and the iterative and data-driven nature of their development.

Good Machine Learning Practice for Medical Device Development: Guiding Principles	
Multi-Disciplinary Expertise Is Leveraged Throughout the Total Product Life Cycle	Good Software Engineering and Security Practices Are Implemented
Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population	Training Data Sets Are Independent of Test Sets
Selected Reference Datasets Are Based Upon Best Available Methods	Model Design Is Tailored to the Available Data and Reflects the Intended Use of the Device
Focus Is Placed on the Performance of the Human-AI Team	Testing Demonstrates Device Performance During Clinically Relevant Conditions
Users Are Provided Clear, Essential Information	Deployed Models Are Monitored for Performance and Re-training Risks are Managed

These 10 guiding principles are intended to lay the foundation for developing Good Machine Learning Practice that addresses the unique nature of these products. They will also help cultivate future growth in this rapidly progressing field.

The 10 guiding principles identify areas where the International Medical Device Regulators Forum (IMDRF), international standards organizations, and other collaborative bodies could work to advance GMLP. Areas of collaboration include research, creating educational tools and resources, international harmonization, and consensus standards, which may help inform regulatory policies and regulatory guidelines.

We envision these guiding principles may be used to:

- Adopt good practices that have been proven in other sectors
- Tailor practices from other sectors so they are applicable to medical technology and the health care sector
- Create new practices specific for medical technology and the health care sector

As the AI/ML medical device field evolves, so too must GMLP best practice and consensus standards. Strong partnerships with our international public health partners will be crucial if we are to empower stakeholders to advance responsible innovations in this area. Thus, we expect this initial collaborative work can inform our broader international engagements, including with the IMDRF.

We welcome your continued feedback through the public docket (FDA-2019-N-1185) at Regulations.gov, and we look forward to engaging with you on these efforts. The Digital Health Center of Excellence is spearheading this work for the FDA. Contact us directly at Digitalhealth@fda.hhs.gov, software@mhra.gov.uk, and mddpolicy-politiquesdim@hc-sc.gc.ca.

U.S. FOOD & DRUG ADMINISTRATION | Health Canada Santé Canada | Medicines & Healthcare products Regulatory Agency

Guiding Principles

1. **Multi-Disciplinary Expertise Is Leveraged Throughout the Total Product Life Cycle:** In-depth understanding of a model's intended integration into clinical workflow, and the desired benefits and associated patient risks, can help ensure that ML-enabled medical devices are safe and effective and address clinically meaningful needs over the lifecycle of the device.

2. **Good Software Engineering and Security Practices Are Implemented:** Model design is implemented with attention to the "fundamentals": good software engineering practices, data quality assurance, data management, and robust cybersecurity practices. These practices include methodical risk management and design process that can appropriately capture and communicate design, implementation, and risk management decisions and rationale, as well as ensure data authenticity and integrity.

3. **Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population:** Data collection protocols should ensure that the relevant characteristics of the intended patient population (for example, in terms of age, gender, sex, race, and ethnicity), use, and measurement inputs are sufficiently represented in a sample of adequate size in the clinical study and training and test datasets, so that results can be reasonably generalized to the population of interest. This is important to manage any bias, promote appropriate and generalizable performance across the intended patient population, assess usability, and identify circumstances where the model may underperform.

4. **Training Data Sets Are Independent of Test Sets:** Training and test datasets are selected and maintained to be appropriately independent of one another. All potential sources of dependence, including patient, data acquisition, and site factors, are considered and addressed to assure independence.

5. **Selected Reference Datasets Are Based Upon Best Available Methods:** Accepted, best available methods for developing a reference dataset (that is, a reference standard) ensure that clinically relevant and well characterized data are collected and the limitations of the reference are understood. If available, accepted reference datasets in model development and testing that promote and demonstrate model robustness and generalizability across the intended patient population are used.

6. **Model Design Is Tailored to the Available Data and Reflects the Intended Use of the Device:** Model design is suited to the available data and supports 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 that the product can safely and effectively achieve its intended use. Considerations include the impact of both global and local performance and uncertainty/variability in the device inputs, outputs, intended patient populations, and clinical use conditions.

7. **Focus Is Placed on the Performance of the Human-AI Team:** Where the model has a "human in the loop," human factors considerations and the human interpretability of the model outputs are addressed with emphasis on the performance of the Human-AI team, rather than just the performance of the model in isolation.

8. **Testing Demonstrates Device Performance During Clinically Relevant Conditions:** Statistically sound test plans are developed and executed to generate clinically relevant device performance information independently of the training data set. Considerations include the intended patient population, important subgroups, clinical environment and use by the Human-AI team, measurement inputs, and potential confounding factors.

9. **Users Are Provided Clear, Essential Information:** Users are provided ready access to clear, contextually relevant information that is appropriate for the intended audience (such as health care providers or patients) including: the product's intended use and indications for use, performance of the model for appropriate subgroups, characteristics of the data used to train and test the model, acceptable inputs, known limitations, user interface interpretation, and clinical workflow integration of the model. Users are also made aware of device modifications and updates from real-world performance monitoring, the basis for decision-making when available, and a means to communicate product concerns to the developer.

10. **Deployed Models Are Monitored for Performance and Re-training Risks Are Managed:** Deployed models have the capability to be monitored in "real world" use with a focus on maintained or improved safety and performance. Additionally, when models are periodically or continually trained 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 as it is used by the Human-AI team.