Good Machine Learning Practice for Medical Device Development: Guiding Principles October 2021 (by FDA, Health Canada and MHRA)

Content current as of: 10/27/2021 (27-Oct-2021)

米国食品医薬品局（FDA）、カナダ保健省（Health Canada）、英国医薬品医療製品規制庁（MHRA）は、共同で Good Machine Learning Practice (GMLP) の開発に情報を提供できる 10 の指針を規定した。これらの指針は、人工知能と機械学習（AI/ML）を使用する上で安全で効果的かつ高品質な医療機器の促進に役立つであろう。

人工知能と機械学習技術は、毎日の医療提供の間に生成される膨大な量のデータから新しく重要な洞察を引き出すことによって、医療を変革する可能性を秘めている。それらはソフトウェアアルゴリズムを使用して実際の使用から学習し、状況によってはこの情報を使用して製品のパフォーマンスを向上させることができる。しかし、その複雑さと、開発の反復的かつデータ駆動的な性質のために、独自の考慮事項が提示されてる。

これらの 10 の指針は、医療機器製品の独自性に対処する Good Machine Learning Practice（GMLP）を制定するための基礎を築くことを目的としている。これらはまた、急速に進歩するこの分野における将来の成長を培うのに有益である。

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

10 の指針は、国際医療機器規制者フォーラム（IMDRF）、国際標準化機関、およびその他の協力機関が GMLP を発展させるために活動できる分野を特定している。協力の分野には、研究、教育ツールとリソースの作成、国際的な調和、および規制政策と規制ガイドラインに情報を提供するのに役立つ可能性のあるコンセンサス基準が含まれる。

我々は、これらの指針が以下の目的に使用されることを想定している。

- 他の部門で証明された優れた実践の採用
- 医療技術やヘルスケア分野に適用できるように、他の分野からのテーラープラクティス
- 医療技術や医療部門に特化した新しい実践の創出

AI/ML 医療機器の分野が発展するにつれて、GMLP のベストプラクティスとコンセンサス標準も必要になる。利害関係者がこの分野で責任あるイノベーションを推進できるようにするためには、国際的な公衆衛生パートナーとの強力なパートナーシップが不可欠である。したがって、この最初の共同作業が、IMDRF を含むより広範な国際的関与に情報を提供できることを期待している。

Regulations.gov のパブリック・ドック（FDA-2019-N-1185）を通じた継続的なフィードバックを歓迎し、これらの取り組みについて関係者と連携することを期待している。Digital Health Center of Excellence は、FDA のためにこの作業の先頭に立っている。Digitalhealth@fda.hhs.gov、software@mhra.gov.uk、および mddpolicy-politiquesdim@hc-sc.gc.ca に直接問い合わせること。

指針原則（Guiding Principles）

1. **総合的な製品ライフサイクルを通じて学際的な専門知識を活用すること:** モデルの臨床ワークフローへの意図された統合、および望ましい利点と関連する患者リスクを深く理解することで、ML 対応の医療機器が安全かつ効果的であり、機器のライフサイクルを通じて臨床的に意味のあるニーズに対応することができる。
2. **優れたソフトウェアエンジニアリングとセキュリティプラクティスを実装すること：** モデル設計は、優れたソフトウェアエンジニアリングのプラクティス、データ品質保証、データ管理、堅牢なサイバーセキュリティのプラクティスなどの「基本」に注意して実装する。これらのプラクティスには、設計、実装、およびリスク管理の決定と根拠を適切に把握して伝達することができ、データの完全性と整合性を保証する、体系的なリスク管理と設計プロセスが含まれる。
3. **臨床試験参加者とデータセットは対象患者の集団を代表すること：** データ収集プロトコルは、対象患者集団の関連する特性(例えば、年齢、性別、性別、人種、民族の観点から)、使用及び測定入力が、臨床試験並びに訓練及び試験データセットにおいて適切なサイズのサンプルで十分に計画され、結果を対象集団に合理的に一般化できることを保証すべきである。これは、あらゆるバイアスを管理し、対象となる患者集団全体で適切かつ一般化可能なパフォーマンスを促進し、有用性を評価し、モデルのパフォーマンスが低下する可能性のある状況を特定するために重要である。
4. **トレーニング（学習）データセットはテストセットから独立すること：** トレーニングデータセットとテストデータセットは、互いに適切に独立しているように選択され、維持する。患者、データ収集、および部位因子を含む全ての潜在的依存源を考慮し、独立性を確保するために対処する。
5. **選択された参照データセットは、利用可能な最良の方法に基づくようにすること:** 参照データセットを開発するための受け入れられた利用可能な最良の方法(すなわち、参照基準)は、臨床的に関連性があり、十分に特徴づけられたデータが収集され、参照の限界が理解されることを保証する。利用可能な場合は、対象となる患者集団全体にわたるモデルの堅牢性と一般化可能性を促進し、実証する、モデル開発とテストにおける受け入れられた参照データセットが使用する。
6. **モデル設計は利用可能なデータに合わせて調整され、医療機器デバイスの使用目的を反映すること:** モデル設計は利用可能なデータに適しており、オーバーフィット、パフォーマンス低下、セキュリティリスクなどの既知のリスクの積極的な軽減をサポートする。本製品に関連する臨床上の有益性とリスクは十分に理解されており、試験のために臨床的に意味のある性能目標を導出するために使用され、本製品が意図された使用を安全かつ効果的に達成できることを支持する。考慮事項には、装置の入力、出力、対象となる患者集団、および臨床使用条件における、全体的および局所的な性能と不確実性/変動性の両方の影響が含まれる。
7. **人間-AI チーム（Human-AI チーム）のパフォーマンスに焦点を置くこと:** モデルが「ループの中の人間（human in the loop）」を持つ場合、人的要因の考慮事項とモデル出力の人間による解釈可能性は、モデルの単独でのパフォーマンスだけでなく、人間-AI チーム（Human-AI チーム）のパフォーマンスに重点を置いて扱われる。
8. **試験は臨床的に適切な状態での医療機器デバイス性能を証明すること:** 統計的に健全な試験計画が開発され、訓練データセットとは独立して臨床的に適切なデバイス性能情報を生成するために実行される。考慮事項には、対象となる患者集団、重要なサブグループ、臨床環境および Human-AI チームによる使用、測定入力、および潜在的交絡因子が含まれる。
9. **ユーザーには、明確で重要な情報を提供すること:** ユーザーには、製品の使用目的と使用の適応症、適切なサブグループに対するモデルのパフォーマンス、モデルのトレーニングとテストに使用されるデータの特性、許容可能な入力、既知の制限、ユーザー・インタフェースの解釈、モデルの臨床ワークフローの統合など、対象ユーザー（医療従事者や患者など）に適した、明確でコンテキストに関連した情報への迅速なアクセスが提供される。また、実際のパフォーマン

ス監視によるデバイスの変更と更新、利用可能な場合の意思決定の基礎、および開発者に製品に関する懸念を伝える手段についてもユーザーに通知する。

10. **導入展開されたモデルのパフォーマンスを監視し、再トレーニングのリスクを管理すること:** 展開されたモデルは、安全性とパフォーマンスの維持または向上に焦点を当て、「現実世界 “real world”」での使用を監視する機能を備える。さらに、モデルが展開後に定期的または継続的にトレーニングされる場合は、Human-AI チームによって使用されるモデルの安全性とパフォーマンスに影響を与える可能性のある、モデルの過剰適合、意図しないバイアス、または劣化(例えば、データセットのドリフト)のリスクを管理するための適切なコントロールを用意する。

(Note: If you plan to take action based on these guidelines, be sure to refer to the original guidelines in English at FDA Web Site.)

https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles

https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles/good-machine-learning-practice-for-medical-device-development-guiding-principles

https://www.fda.gov/media/153486/download
https://hiropharmaconsulting.com/news/2021/10/30/vol-3-no-8/

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.