「医療AI開発・使用における10の原則」について

JCR人工知能診療委員会

このたび，米英加による共同声明「医療AI開発・使用における10の原則」が発表された。
https://www.gov.uk/government/publications/good-machine-learning-practice-for-medical-device-development-guiding-principles

日本語訳に加えて，理解を助けるため具体例（良い例・悪い例を1つずつ）を挙げた。

「医療AI開発・使用における10の原則」作成の目的

冒頭部分の日本語訳

米国食品医薬品局（FDA），カナダ保健省，英国医薬品・ヘルスケア製品規制庁（MHRA）は，Good Machine Learning Practice（GMLP）の策定に役立つ10の原則を共同で作成した。これらの指針は，人工知能・機械学習を用いた安全かつ高品質な製品の開発に役立つ。

人工知能・機械学習技術には大きな力があり，医療を変革する可能性を秘めているが，ソフトウェアの複雑さとデータへの強い依存性から，特有の懸念事項がある。

10の原則は，国際医療機器規制フォーラム（IMDRF），国際標準化団体，学会等が，GMLPを推進するために取り組むべき内容を示している。研究，教育ツールや教育資料の作成，国際ハーモニゼーション，規制政策などの作成に役立つ。

解　説

臨床研究や治験等に関わると，Good Clinical Practice（GCP：医薬品の臨床試験の実施基準），Good Manufacturing Practice（GMP：医薬品等の製造管理・品質管理の基準），Good Laboratory Practice（GLP：医薬品の安全性に関する非臨床試験の実施基準）といった基準を目にすることがある。この文脈で，機械学習を用いた製品の開発における実施基準として，Good Machine Learning Practice（GMLP）がこのたび提案された。

10の原則の各項目を端的に述べると，

1．製品のライフサイクル全体を通して，多分野の専門家が参加する。
2．優れたソフトウェア工学とセキュリティ保証が行われる。
3．臨床試験の参加者及びデータセットが対象となる患者集団を代表している。
4．トレーニング用のデータセットがテスト用のデータセットから独立している。
5．選択された基準データセットが実現可能な最善の方法に基づいて作られている。
6．モデルは，入手可能なデータと機器の使用目的に合わせて作成される。
7．人間とAIがチームとなった状態のパフォーマンスを評価する。
8．臨床的に適切な条件でデバイス（AIモデル）の性能を実証テストする。
9．ユーザーは明確で重要な情報を提供される。
10．インストールされたモデルの性能を検証し，再トレーニングのリスクを管理する。

である。

機械学習の研究者にとっては当たり前の原則と思われるかもしれないが，これらが明文化されたことで，国際機関や各学会から，GMLPに準拠した基準が今後発表されることが予想される。

以下で内容を詳しく見ていこう。

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.

日本語訳

原則1　製品のライフサイクル全体を通して，多分野の専門家が参加する

臨床ワークフローへのモデルの組み込み，期待される利益とそれに伴う患者のリスクを深く

理解することで，機器のライフサイクルを通じて，機械学習対応の医療機器が安全で効果的であり，臨床的に意味のあるニーズに対応することが保証される。

良い例

内科医，画像診断医，データサイエンティスト，倫理・法律の専門家，企業からなる多分野専門家がチームを組んで1つの製品の開発，メンテナンスに取り組む。

悪い例

画像診断医不在のチームが画像診断AI製品を開発する。(理由：このチームは，適切な教師データが作成できなかったり，患者のリスク・ベネフィットを正しく評価できなかったりする可能性があり，チームの構成としては不十分である。)

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.

日本語訳

原則2　優れたソフトウェア工学とセキュリティ保証が行われる

モデル設計は，優れたソフトウェア工学，データ品質保証，データ管理，および強固なサイバーセキュリティという「基本」を念頭に置いて実施され，設計，実装，リスクマネージメントに関する決定とその根拠を適切に把握・伝達し，データの信正性と完全性を確保することができる体系的なリスク管理と設計プロセスが含まれる。

良い例

ソフトウェア工学，サイバーセキュリティに十分な知識をもつエンジニアが開発チームに参加し，予期していない入力に対するAIの挙動を十分に検討している。

悪い例

サイバーセキュリティの知識が不十分な状態で製品を開発する。(理由：このような製品は，安全な環境では動作するかもしれないが，悪意の有無に関わらず予定外の入力に対して不安定な挙動を示す可能性がある。これでは臨床使用するためには不十分である。)

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.

日本語訳

原則3　臨床試験の参加者及びデータセットが対象となる意図した患者集団を代表している

対象とする患者群に適切に使用できるモデルを作成するため，対象とする患者集団の特性（年齢，性別，人種，民族など），使用方法，および測定入力が，十分な臨床試験，トレーニング及びテストデータセットのサンプル数とともにデータ収集のプロトコルに示されていること。これは，バイアスを管理し，対象となる患者集団における汎化性能を向上させ，有用性を評価し，モデルの性能が低下する可能性がある状況を特定するために重要である。

良い例

開発準備段階において，AI製品が使用される状況（対象患者，対象疾患）を十分に検討したうえで，適切な教師データを集める。

悪い例

アメリカの非喫煙者のデータベースを利用して，日本人の喫煙者の疾患を診断するAIを開発する。(理由：開発に使用した患者集団と実用時の患者集団が異なっているため，意図した性能を発揮できない。)

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.

日本語訳

原則4　トレーニング用のデータセットがテスト用のデータセットから独立している

トレーニングデータセットとテストデータセットは，互いに独立するように選択・維持されていること。独立性を確保するために，患者，データ取得，施設の要因など，依存性の原因となり得るものはすべて考慮し，対処する。

良い例

A, B, C, Dの4施設から臨床データを集めたので，A, Bの600例を教師データセットに，C, Dの400例をテストデータセットとしてあらかじめ分割した。

悪い例

トレーニングに用いたデータセットと同一のデータセットを使用して性能を評価する。(理由：これを行うと，過学習によっていくらでも性能を高められるが，未知のデータに対する性能（汎化性能）がまったく評価できない。)

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.

日本語訳

原則5　選択された基準データセットが実現可能な最善の方法に基づいて作られている

（教師データとしての）基準データセットはが広く一般的に受け入れられたるための最善の方法は，臨床的に適切に確定診断されたデータが収集され，また，基準データセットの限界が理解されていることである。可能であれば，モデルの開発と試験において，モデルの頑健性と対象となる患者集団における一般化を促進し実証する，受け入れられた基準データセットを使用する。

良い例

GCP や ICH（　＝International Council for Harmonisation）などに沿った質の高い臨床研究プロトコルを作成し，これに基づいて収集された臨床データを性能評価のデータセットとして使用する。また，どのような患者集団が含まれているか（含まれていないか）が，使用者にわかるように整理されている。

悪い例

倫理委員会の承認を経ていない方法で収集された臨床データを用いて，性能評価を行う。(理由：医の倫理の観点で問題であるだけでなく，データの真正性も疑わしく，性能評価の結果を信頼することができない。)

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.

日本語訳

原則6　モデルは，入手可能なデータと機器の使用目的に合わせて作成される

モデルは，入手可能なデータに適合するように，またオーバーフィッティング，性能劣化，セキュリティリスクなどの既知のリスクを積極的に軽減するように設計されていること。製品に関連する臨床上の利点とリスクは臨床的に意味のある性能目標を導き出すために，また製品が意図した用途を安全かつ効果的に達成できるために十分に理解されていること。考慮すべき点は，グローバルおよびローカルな性能の影響や，機器の入力，出力，対象となる患者集団，臨床使用条件における不確実性／変動性などである。

良い例

疾患Aは有病率が低く，教師データとして10症例しか集まらなかった。パラメーター数の多いAIモデルの開発は難しいと考えられた。そこで，もともと使用予定であったAIモデルから変更し，少ない症例数でも使用できるとされるAIモデルを使用することにした。

悪い例

ある論文の中で，あるモデル（ニューラルネットワーク）が非常に性能を示したと報告されていたので，この論文を読んだ研究者が，十分な検討をせずに，そのモデルを自分の手持ちのデータに使用してAI製品を開発した。（理由：利用可能な数多くのAIアルゴリズム・AIモデルのなかから，手持ちのデータに一番ふさわしいものを選択すべきである。）

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.

日本語訳

原則7　人間とAIがチームとなった状態のパフォーマンスを評価する

意思決定にAIだけでなく人間が参加している場合は，モデル単体の性能だけでなく，Human-AIチームの性能に重点を置いて，ヒューマンファクターの検討や，モデルの出力を人間がどのように解釈するかに焦点が当てられる。

良い例

CTで肺結節を検出するAIモデルの性能を評価する際に，実際に使用される場面を考慮して，放射線科医がAIを使用した状態での性能（感度，特異度など）を評価した。

悪い例

あるAI製品を開発している。実臨床では放射線科医がAIの提案結果を考慮しながら診断していく使い方を想定しているが，性能評価のときに協力してくれる放射線科医がいなかったため，AI単独での性能（感度，特異度）だけを測定して結論とした。（理由：実際に放射線科医がAIを使用したときの性能が不明である）

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.

日本語訳

原則8　臨床的に適切な条件でデバイス（AIモデル）の性能をテストする

教師データセットとは別なデータを用いて，AIの性能情報を生成するために，統計学的に健全なテストが実施されること。想定される患者集団，重要なサブグループ，臨床環境，Human-AIチームによる使用，測定入力，および潜在的な交絡因子などを考慮する。

良い例

40-80歳台の肺CTの診断を目的としたAIモデルを作成した。性能評価のために，教師データでは使用しなかった症例で，なおかつ40-80歳台の肺CTを用いて，性能評価を行った。その際に，統計学者に加わってもらい，統計学的に問題がないことを検証してもらった。

悪い例

日本人から収集したデータをもとに，日本人の診断を目的としたAIモデルを開発した。このモデルの性能を評価したいが，ちょうどよいテストデータセットが入手できなかったため，アメリカ人から収集したデータをテストデータとして使用し性能を評価した。（理由：評価時と実用時の患者集団が異なるため，実用時の性能を評価しているとはいえない）

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.

日本語訳

原則9　ユーザーは明確で重要な情報を提供される

ユーザーは，製品の使用目的と適応症，適切なサブグループに対するモデルの性能，モデルのトレーニングとテストに使用されたデータの特徴，許容できる入力，既知の制限，ユーザーインターフェースの解釈，モデルの臨床ワークフローへの統合など，意図された対象者（医療提供者や患者など）に適した明確な情報にすぐにアクセスできること。また，ユーザーは，機器の修正や実臨床での性能モニタリングによる最新情報，利用可能な場合は意思決定の根拠，製品に関する懸念事項を開発者に伝える手段についても知らされなければならない。

良い例

働いている病院にAI製品がインストールされた。提供元の企業から，どういった場面でAIを使ってよいか，逆にどういった場面ではAIを使えないかについての十分な説明を受けた。

悪い例

ある日，放射線科医として働いている病院に出勤すると，読影端末にAI製品がインストールされていた。使用方法や使用条件については何も知らされていないが，起動してみたところ，なにやら画面にAIによる診断が出てきた。よく理解できていないが，AIはそもそもブラックボックスであるから，細かいことは気にせず使っている。（理由：想定された方法とは異なる使い方によって深刻な結果を招く可能性がある。）

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.

日本語訳

原則10　配備されたモデルの性能をモニターし，再トレーニングのリスクを管理する

配備されたモデルは，安全性や性能の維持・向上に重点を置いて，実臨床での使用状況をモニタリングすることができる。また，配備後にモデルを定期的または継続的にトレーニングする場合，Human-AIチームが使用するモデルの安全性や性能に影響を与える可能性のあるオーバーフィット，意図しないバイアス，モデルの劣化（データセットのドリフト（トレーニング時と実用時のデータのずれ）など）のリスクを管理するための適切なコントロールが必要である。

良い例

AI製品を開発し販売を開始した。10箇所の病院にインストールして使用してもらっている。定期的に各病院を訪問し，実際のデータにおける診断精度をチェックしている。病院の先生からは，説明が多いと厄介がられることもあるが，性能を保証するためには重要なことであるから継続している。

悪い例

ある放射線科医が使用しているAIが，稀な疾患の診断において誤った診断名を提示した。この症例を教師データとして再トレーニングしたほうが，AIは賢くなるはずであるから，実際に再トレーニングしてみた。再トレーニング後は性能が変化しているはずであるが，稀な疾患を加えたのだから性能は改善しているはずだ。性能評価には時間も手間もかかるので，それは省略して，さっそく実臨床に使用する。（理由：再トレーニングによって性能が低下することがあるので，性能評価を経ずに実臨床に使用することは認められない。）