【報告様式Ａ】

課題管理番号:24mk010101240j0003　作成／更新日:令和 7 年 5 月 30 日

日本医療研究開発機構　　医薬品等規制調和・評価研究事業
事後評価報告書

公開

Ⅰ 基本情報

研究開発課題名：　（日本語）AI を利用した医療機器プログラムの市販後学習時の性能評価に資する研究

（英　語）Study on relevant performance evaluating process after post-marketing training of software as medical devices utilizing artificial intelligence with machine learning

研究開発実施期間：令和 4 年 7 月 15 日～令和 7 年 3 月 31 日

研究開発代表者　氏名：(日本語) 中岡　竜介

(英　語) Ryusuke Nakaoka

研究開発代表者　所属機関・部署・役職：

（日本語）国立医薬品食品衛生研究所　医療機器部　室長

（英　語）Section Head, Division of Medical Devices, National Institute of Health Sciences

II 研究開発の概要

研究開発の成果およびその意義等

和文：

近年、機械学習に基づいた AI を利用した様々な医療機器プログラムが開発されており、主に画像診断支援を目的とした製品の承認が増えつつある。当該プログラムの研究開発は世界中で活発に行われており、米国においては 2018 年に複数の製品が FDA により承認されているが、近年着目されている深層学習の特性をその主機能に活用した製品の承認数は明確にされておらず、その正確な数は未知数である。しかしながら、本邦での承認数はまだ少数であることから、その研究開発を促進するため、AI 利用医療機器プログラムの市販後学習による性能向上を速やかに承認するための法体系が整備された。

当該プログラムの性能変化に係る検証は QMS の設計開発関連手順により一部代用できることとなったが、その実運用においては、市販後学習プロセスとその特性に応じた性能評価プロセスとを連動させる必要がある。過去の市販後学習プロセスに関する研究により、市販後学習に使用するデータのアノテーション最適化方法は一律に規定できず当該プログラムの臨床的位置付けに応じて変更する必要性があること、臨床的位置付けに応じたデータ特性や収集対象となる母集団が重要だが意図せず市販後学習時の学習データが承認時のテストデータと性質が異なってしまう可能性があること、結果として実施された性能評価における信頼性が問題となり得ること等を示すことができた。しかしながら、変更計画確認申請制度（IDATEN）の実運用において、もう一方の柱となる市販後学習時の性能評価プロセスに資する研究は未実施であったことから、市販後学習時の医療機器プログラムに対する適切な性能評価プロセス構築の一助となる成果が必要であり、そのための研究の実施が求められていた。

当該プログラムの性能評価においては、最初の承認申請時に使用したテストデータを用いて比較検証することが想定されるが、同一のテストデータを用いた検証を何度も繰り返した場合、見かけ上、良い結果が得られたとしても、選定したモデル自体の性能は母集団に対して最適化されていない可能性があるため、十分な数のテストデータから必要数をランダムに抽出して使用する等、真値からのバイアスを補正するための対策が必要になることが想定される。また、AI を利用する医療機器（CT、MRI 等）の性能が向上した場合、市販後学習に使用するデータと承認申請時に使用したテストデータとの間で性質が異なることになるため性能評価の妥当性を慎重に検討する必要が生じる等、新たな問題が生じることが予想される。よって、AI の市販後学習に伴う性能変化を適切に評価するためには、想定される問題点を把握した上で、性能評価に使用するテストデータに関する要件整理が必要であった。国際的にも当該プログラムの性能評価の重要性は着目されており、ISO/IEC で進みつつある AI 利用医療機器プログラムに関連する国際標準作成作業において国内における開発状況も踏まえた内容を本邦から提案できるよう、科学的根拠を積み上げておく必要性も生じている。

そこで、本研究事業では、近年開発されている機械学習等の AI アルゴリズム特性である学習データ追加に伴う性能変化・向上（市販後学習）機能を十分に活かした、AI 利用医療機器プログラムの迅速な社会実装に必要な規制環境の整備に資する結果を得ることを目的に、その調査研究に加え、既承認品を模した AI 利用モデルプログラムを使用した実証実験を実施した。実施項目毎に概要を記載する。

１）研究総括及びガイダンス案作成

ガイダンス案作成において参考となる海外、特に米国における ML 利用医療機器の性能評価に関する現状と薬事上の要求事項に関する調査を行い、米国でも利用者による市販後学習が可能な ML 利用医療機器の流通を未だに認めておらず製造販売業者のみが実施可能であること、医療機器プログラムの円滑なアップデートに資する施策を構築中であること等が判明した。後者の施策は本邦で導入済みの施策と

ほぼ同じ内容であったことから、本研究で得られた成果を基にガイダンス案を作成すれば米国に遅れをとることはないことが予想された。そこで、構築した連携体制を利用して、日本医療機器産業連合会（医機連）の関連ワーキンググループに研究で得られた成果を提供し、その活用を討議してもらったが、重要性は理解されたものの、種々のモダリティを利用した研究では予備成果しか得られていないことからその一般化には今後の継続的な研究が必要であること、限られたデータ数の活用法に関する研究を深掘りしてほしいこと等のコメントが寄せられた。また、規制当局からも同様の指摘があったことから、関連研究のさらなる推進とその成果を基にした素案の継続的なブラッシュアップが必須との結論になった。

２）市販後学習時の性能変化に影響する因子の探索

既承認品を模した AI 利用モデル医療機器プログラムを用いた検証により、市販後学習時の性能変化に影響する因子をいくつか特定することができた。さらに、入力画像や目的が異なる AI 利用医療機器プログラムの場合にも類似の因子が影響を及ぼす可能性についての予備検討を追加実施した。その結果、モダリティの種類にかかわらず、類似の一部因子が性能に影響すること、市販後学習に使用するデータの種類に応じて破滅的忘却が生じて再学習前のテストデータに対する性能が劣化することが認められた。

３）市販後学習時の適切な性能評価系の探索と妥当性検証

既存の製品を模した AI 利用モデルプログラム医療機器を利用して、市販後学習を実施した医療機器プログラムのテストデータ再利用による性能へのバイアス混入問題とその対策を検討した結果、アルゴリズムの設計に市販後テストデータによる結果を繰り返し利用することでバイアスが混入することが確認された。一方、そのバイアスが緩和・補正できる手法が存在することも確認できた。

また、市販前と市販後とで学習データの基となる患者群が異なると、２）の検討で確認された事象と同様、市販後学習により得られたモデルを市販前データと同等の患者群に適用すると性能が低下することが示唆されたことから、市販後学習後の性能評価においては、使用した学習データ及びテストデータに何らかの要件を設けるか、バイアス軽減等の対策を検討する必要があると考えられた。

Real world データに適用した場合の underspecification の問題についても検討を行ったところ、underspecification の問題が起こった場合に増えるとされるモデルのテスト性能のばらつきは市販後学習によって小さくなったことから、適切な市販後学習プロセスの構築により underspecification の問題が緩和されることが示唆された。また、real world におけるデータとの類似性を考慮した加工テストデータを利用する stress test により，real world の未知データに対する性能の振る舞いを予測できる可能性が示唆された。

実証研究を通して、特定された市販後学習時の性能変化に影響する因子については、その一部が適用するモダリティの種類に関係なく影響を及ぼすこと、モダリティの性質に応じた因子が存在しうることが判明したことから、市販後学習の種類・方法や目的に応じた特異的な留意点の抽出が、一般的な留意点の抽出と並行して必要であり、各々をカバーしたガイダンス類の必要性を明らかにした。加えて、その学習モデルによっては、市販後学習後の性能を評価する上でのテストデータ再利用が不適である可能性が判明したことから、再使用可能なケースの分類がガイダンス類の作成に必要であることが判明した。一方、業界及び規制当局から指摘があったように、本事業で得られた成果を提言等に作成するためにはさらなる成果の積み上げが必要であることから、産官学の協力を得た上で本研究を継続することが、本邦における AI 利用医療機器プログラムの迅速な社会実装とその発展に不可欠であると考えられる。

英文：

Recently, various medical device programs (software as a medical device: SaMD) utilizing artificial intelligence (AI)-based on machine learning have been developed, and the many SaMDs, especially for image diagnosis support, have been approved for their marketing. In the United States, many AI-based SaMDs have been approved by the FDA since 2018; however, their real approved number that fully leverage the characteristics of deep learning for their intended use remains unclear but may be limited. To promote their research and development, Japan established a legal framework by expediting the approval of performance improvements through their post-market training.

In actual implementation, it is necessary to link the post-market training process with performance evaluation processes tailored to the characteristics of the post-market training process. Previous our research on post-market training processes has shown that there is no uniform method for optimizing the annotation of data used in post-market training, and that it is necessary to modify the method according to their actual clinical usage. Additionally, while data characteristics and the target population for data collection are important depending on the clinical usage, there is a risk that the data used in post-market training may unintentionally differ in nature from the test data used at the time of market approval process, which could compromise the reliability of the performance evaluation. However, it is also necessary to perform research contributing to the performance evaluation process during post-marketing training to develop adequate process for accepting post-marketing training for AI based SaMD.

Therefore, this research has been done to obtain results that contribute to the establishment of a regulatory environment necessary for the rapid social implementation of AI-based SaMDs by fully utilizing the performance changes and improvements (post-market training) associated with the addition of training data, which is a characteristic of machine learning algorithms that have been developed in recent years. In addition to conducting investigative research on AI-based SaMDs, we conducted simulations using mock AI-based SaMDs mimicking real market approved SaMDs.

As results of the simulations, several factors which influence performance changes via post-market training of the mock AI-based SaMD have been identified, and some of them may influence them regardless of medical image type. It has been also revealed that a deterioration in their performance on test data prior to post-market learning occurs by “catastrophic forgetting” depending on the type of data used. In addition, we have found that bias may be introduced in SaMD using some type of machine learning algorithm by the repeated-use of test data for evaluating performance after its post-market training by another simulation. We have also found a possible method to ease an effect of “underspecification”, which is useful to design appropriate process for evaluate actual performance of AI-based SaMD.

However, discussion in collaboration with industry, regulation, and academic members has revealed necessity of more mature results utilizing various type of AI-based SaMDs if possible, in order to develop guidance how to perform post-market training including appropriate performance evaluation process for AI-based SaMD, suggesting huge demand for continuing the research and acquiring further knowledges for drafting the guidance for rapid social implementation and development of AI-based SaMDs.