DOI: 10.1002/hcs2.114

COMMENTARY

HEALTH CARE SCIENCE

Toward real-world deployment of machine learning for health care: External validation, continual monitoring, and randomized clinical trials

Han Yuan

Centre for Quantitative Medicine, Duke-NUS Medical School, Singapore, Singapore

Correspondence
Han Yuan, Centre for Quantitative Medicine, Duke-NUS Medical School, 8 College Rd, Singapore 169857, Singapore.
Email: yuan.han@u.duke.nus.edu

KEYWORDS
machine learning, real-world deployment, external validation, continual monitoring, randomized clinical trials

Funding information
None

1 | OVERVIEW

Machine learning (ML) has been increasingly used for tackling various diagnostic, therapeutic, and prognostic tasks owing to its capability to learn and reason without explicit programming [1]. Most developed ML models have had their accuracy proven through internal validation using retrospective data. However, external validation using retrospective data, continual monitoring using prospective data, and randomized controlled trials (RCTs) using prospective data are important for the translation of ML models into real-world clinical practice [2]. Furthermore, ethics and fairness across subpopulations should be considered throughout these evaluations.

2 | EXTERNAL VALIDATION

Different from internal validation, which evaluates the performance of ML using a subset of the original datasets, external validation assesses ML models in contexts that may vary subtly or considerably from the one in which they were developed [3]. External validation serves to rectify inflated estimates of ML capabilities owing to overfitting and guarantees the generalizability and transportability of ML models across diverse populations [4]. For external validation, researchers can leverage the abundant resources of publicly accessible databases such as PhysioNet [5]. Three external validation scenarios are recommended after identifying a suitable database with a sufficient sample size to guarantee testing robustness [6]. The first involves directly deploying the trained ML models on external data to simulate a brand-new scenario without previous data [6]. The second entails using a large training data set from the new scenario to fine-tune the developed models, simulating that ample data have been collected in the external context [7]. The third scenario represents an intermediate situation wherein new data are gradually fed into the ML models to simulate a scenario where the models are deployed in a new setting, new data are incrementally collected, and the models are updated iteratively with the newly collected data [8]. Most existing studies have focused on the direct deployment of ML models for diagnostic, therapeutic, and prognostic tasks [9]. Holsbeke et al. [10] deployed previously published diagnostic ML models for detecting adnexal mass malignancy across multiple medical centers in different countries

Abbreviations: FDA, US Food and Drug Administration; ML, machine learning; RCTs, randomized controlled trials.

with different population characteristics. For external validation of therapeutic ML models, a pertinent reference is a study investigating the survival benefits of adjuvant therapy in breast cancer where researchers evaluated ML models, which were originally developed using populations from the United Kingdom, in clinical settings in the United States [11]. In the realm of prognostic tasks, Clift et al. [12] offered a comprehensive approach to externally validate ML models in the context of predicting the 10-year risk of breast cancer-related mortality, detailing methods for sample size calculation, population identification, outcome definition, and performance evaluation. In addition to assessing model performance, similarity between the original training datasets and external validation datasets can be quantified to enable the elucidation of performance degradation and further identify potential avenues for model enhancement [13].

3 | CONTINUAL MONITORING

Following large-scale external validation using retrospective data, the subsequent step in implementation is prospective evaluation in the specific setting where an ML model is to be deployed [14]. Specifically, ML models receive prospective data, make predictions accordingly, and are evaluated within a predefined time frame [14]. Compared with the first step, continual monitoring is used to identify data distribution drift, control model quality, and trigger system alarms when an ML model deviates from its normal behavior in the target setting [15]. Because the operation and monitoring of ML models are mainly conducted by clinical professionals, developers should focus on translation of the developed ML models into a user-friendly clinical practice. The first aspect is the operation of ML models in an offline hospital system where allocated computation resources would be limited for low latency in responding to other functions inside the system. The second aspect is the development of a secure and privacy-aware maintenance method for quickly addressing potential technical collapses while minimizing direct access to patients' private data. The last aspect is the development of a user-friendly interface such as an Android app [16] or web-based software [17] that facilitates the use of ML models by health care professionals and comprehends their suggestions. It should be emphasized that the application of ML in a prospective clinical setting should be designed to operate independently from, and not interfere with, existing clinical decision-making processes. This precaution is necessary to avoid any potential adverse impact on the existing health care quality. Exemplary continual monitoring of therapeutic ML models can be seen in the work of Wissel et al. [18]. Those authors conducted a prospective, real-time assessment of ML-based classifiers for epilepsy surgery candidacy at Cincinnati Children's Hospital Medical Center. To mitigate any risks associated with ML classifiers, patients who were deemed appropriate surgical candidates by the algorithm were subjected to manual review by two expert epileptologists, with final decisions on their surgical candidacy confirmed via a comprehensive expert chart review. A critical insight from the study was that effective monitoring necessitates a synergistic collaboration between clinicians, who provide essential medical expertize, and information technology professionals, who contribute research and operational knowledge [19, 20]. Assuming that an ML tool demonstrates accurate prospective diagnostic capabilities in the target setting, its developers should pursue approval for further RCTs from administrative ethics committees.

4 | RANDOMIZED CONTROLLED TRIALS

The last step toward the real-world implementation of ML tools is classic four-phase RCTs. To ensure safety in real-life scenarios, absolutely ML-based interventions are likely to be avoided. We recommend designing RCTs to compare the accuracy and diagnosis time for clinicians with ML models (intervention group) and without ML models (control group) [21–23]. For instance, He et al. [24] implemented RCTs to demonstrate that ML-guided workflows reduced the time required for sonographers and cardiologists in the diagnoses of left ventricular ejection fraction. Specifically, the first step in RCTs is to seek ethical approval from an institutional review board to ensure that the RCTs comply with ethical standards and regulations. Then, researchers can proceed with Phase I of the clinical trial to assess safety (whether the introduction of an ML model distracts clinicians and impairs their diagnoses) and to identify specific scenarios in which ML should be used. In Phase II, a few hundred patients are recruited to assess whether statistically significant improvements result from the use of ML tools in clinicians' diagnoses. In Phase III, several hundred or even several thousand patients are recruited to validate the safety and effectiveness of the ML tool, demonstrating its superiority over other existing solutions. If the ML tool receives approval from the administrative agency after Phase III, researchers can then investigate its effectiveness and safety in a wider range of patients in Phase IV. Upon demonstrating efficacy through rigorously conducted RCTs, ML tools can receive approval from national regulatory agencies such as the US Food and Drug Administration (FDA)

for commercialization [25]. A paradigmatic illustration of RCTs for diagnostic ML models can be found in the research by Titano et al. [26]. Those authors developed three-dimensional convolutional neural networks to diagnose acute neurological events using head computed tomography images. The efficacy and efficiency of ML models were subsequently validated in a randomized, double-blind, prospective trial. For therapeutic ML models, we suggest referring to Nimri et al. [27]. The researchers conducted multicenter and multinational RCTs to compare ML with physicians from specialized academic diabetes centers in optimizing insulin pump doses. In the realm of prognostic ML models, researchers from the Mayo Clinic implemented RCTs to assess the effectiveness and efficiency of ML models in predicting 1-year occurrence of asthma exacerbation [28]. A detailed guideline for conducting RCTs on ML for health care could benefit from the FDA's Policy for Device Software Functions and Mobile Medical Applications [29], which includes specific provisions for medical applications that apply ML algorithms [30].

5 | TOWARD REAL-WORLD DEPLOYMENT

Alongside population-level evaluations, there has been burgeoning awareness about the ethical implications of ML models, which have been revealed to diagnose, treat, and bill patients inconsistently across subpopulations [31]. Therefore, it is imperative to ensure equity of patient outcomes, model performance, and resource allocation across subpopulations in the real-world deployment of ML models [31–33]. Thompson et al. [34] proposed a reference framework to mitigate ML biases using two recalibration modules. The first module adjusted the decision cutoff threshold for subpopulations affected by bias, and the second recalibrated model outputs, enhancing their congruence with the observed events. Chen et al. [31] systematically summarized the path toward deployment of ethical and fair ML in medicine, which includes a diverse subpopulation collection using federated learning, fairness principles, operationalization across health care ecosystems, and independent regularization and governance of data and models to avoid disparities. Apart from various performance assessments, clinicians' endorsement and patients' approval of ML models should be thoroughly integrated into the evaluation processes [31, 35].

In this commentary, we elucidate three indispensable evaluation steps toward the real-world deployment of ML within the health care sector and provide examples of diagnostic, therapeutic, and prognostic tasks. In light of these, we encourage researchers to move beyond retrospective and within-sample validation and toward the practical implementation at the bedside rather than leaving developed ML models buried within the archived literature.

AUTHOR CONTRIBUTIONS

Han Yuan: Conceptualization (lead); data curation (lead); formal analysis (lead); investigation (lead); methodology (lead); writing—original draft (lead); writing—review and editing (lead).

ACKNOWLEDGMENTS

I would like to acknowledge Prof. Nan Liu at Duke-NUS Medical School for his invaluable support.

CONFLICT OF INTEREST STATEMENT

The author declares no conflict of interest.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

ETHICS STATEMENT

This study is exempt from review by the ethics committee because it did not involve human participants, animal subjects, or sensitive data collection.

INFORMED CONSENT

Not applicable.

ORCID

Han Yuan http://orcid.org/0000-0002-2674-6068

REFERENCES

1. Ben-Israel D, Jacobs WB, Casha S, Lang S, Ryu WHA, de Lotbiniere-Bassett M, et al. The impact of machine learning on patient care: a systematic review. Artif Intell Med. 2020;103:101785. https://doi.org/10.1016/j.artmed.2019.101785
2. Triantafyllidis AK, Tsanas A. Applications of machine learning in real-life digital health interventions: review of the literature. J Med Internet Res. 2019;21(4):e12286. https://doi.org/10.2196/12286
3. Cabitza F, Campagner A, Soares F, García de guadiana-Romualdo L, Challa F, Sulejmani A, et al. The importance of being external. methodological insights for the external validation of machine learning models in medicine. Comput Methods Programs Biomed. 2021;208:106288. https://doi.org/10.1016/j.cmpb.2021.106288
4. Ho SY, Phua K, Wong L, Bin Goh WW. Extensions of the external validation for checking learned model interpretability and generalizability. Patterns. 2020;1(8):100129. https://doi.org/10.1016/j.patter.2020.100129
5. Moody GB, Mark RG, Goldberger AL. PhysioNet: a web-based resource for the study of physiologic signals. IEEE Eng Med Biol Mag. 2001;20(3):70–5. https://doi.org/10.1109/51.932728

6. Alsentzer E, Rasmussen MJ, Fontoura R, Cull AL, Beaulieu-Jones B, Gray KJ, et al. Zero-shot interpretable phenotyping of postpartum hemorrhage using large language models. npj Digit Med. 2023;6:212. https://doi.org/10.1038/s41746-023-00957-x
7. Yang Z, Mitra A, Liu W, Berlowitz D, Yu H. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. Nat Commun. 2023;14(1):7857. https://doi.org/10.1038/s41467-023-43715-z
8. Soltoggio A, Ben-Iwhiwhu E, Braverman V, Eaton E, Epstein B, Ge Y, et al. A collective AI via lifelong learning and sharing at the edge. Nat Mach Intell. 2024;6:251–64. https://doi.org/10.1038/s42256-024-00800-2
9. Yuan H, Yu K, Xie F, Liu M, Sun S. Automated machine learning with interpretation: a systematic review of methodologies and applications in healthcare. Med Adv. 2024;2(3):1–33. https://doi.org/10.1002/med4.75
10. Van Holsbeke C, Van Calster B, Bourne T, Ajossa S, Testa AC, Guerriero S, et al. External validation of diagnostic models to estimate the risk of malignancy in adnexal masses. Clin Cancer Res. 2012;18(3):815–25. https://doi.org/10.1158/1078-0432.CCR-11-0879
11. Alaa AM, Gurdasani D, Harris AL, Rashbass J, van der Schaar M. Machine learning to guide the use of adjuvant therapies for breast cancer. Nat Mach Intell. 2021;3:716–26. https://doi.org/10.1038/s42256-021-00353-8
12. Clift AK, Dodwell D, Lord S, Petrou S, Brady M, Collins GS, et al. Development and internal-external validation of statistical and machine learning models for breast cancer prognostication: cohort study. BMJ. 2023;381:e073800. https://doi.org/10.1136/bmj-2022-073800
13. Kouw WM, Loog M. A review of domain adaptation without target labels. IEEE Trans Pattern Anal Mach Intell. 2021;43(3):766–85. https://doi.org/10.1109/TPAMI.2019.2945942
14. Akhlaghi H, Freeman S, Vari C, McKenna B, Braitberg G, Karro J, et al. Machine learning in clinical practice: evaluation of an artificial intelligence tool after implementation. Emer Med Australas. 2024;36(1):118–24. https://doi.org/10.1111/1742-6723.14325
15. Paleyes A, Urma RG, Lawrence ND. Challenges in deploying machine learning: a survey of case studies. ACM Comput Surv. 2023;55(6):1–29. https://doi.org/10.1145/3533378
16. Kumar N, Narayan Das N, Gupta D, Gupta K, Bindra J. Efficient automated disease diagnosis using machine learning models. J Healthc Eng. 2021;2021:9983652. https://doi.org/10.1155/2021/9983652
17. Imrie F, Cebere B, McKinney EF, van der Schaar M. AutoPrognosis 2.0: democratizing diagnostic and prognostic modeling in healthcare with automated machine learning. PLOS Digital Health. 2023;2(6):e0000276. https://doi.org/10.1371/journal.pdig.0000276
18. Wissel BD, Greiner HM, Glauser TA, Holland-Bouley KD, Mangano FT, Santel D, et al. Prospective validation of a machine learning model that uses provider notes to identify candidates for resective epilepsy surgery. Epilepsia. 2020;61(1):39–48. https://doi.org/10.1111/epi.16398
19. Kanbar LJ, Wissel B, Ni Y, Pajor N, Glauser T, Pestian J, et al. Implementation of machine learning pipelines for clinical practice: development and validation study. JMIR Med Inform. 2022;10(12):e37833. https://doi.org/10.2196/37833
20. Yuan H, Hong C, Jiang PT, Zhao G, Tran NTA, Xu X, et al. Clinical domain knowledge-derived template improves post hoc AI explanations in pneumothorax classification. J Biomed Inf. 2024;156:104673. https://doi.org/10.1016/j.jbi.2024.104673
21. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. Lancet Oncol. 2019;20(5):e262–73. https://doi.org/10.1016/S1470-2045(19)30149-4
22. Seneviratne MG, Shah NH, Chu L. Bridging the implementation gap of machine learning in healthcare. BMJ Innovations. 2020;6(2):45–7. https://doi.org/10.1136/bmjinnov-2019-000359
23. Tricco AC, Hezam A, Parker A, Nincic V, Harris C, Fennelly O, et al. Implemented machine learning tools to inform decision-making for patient care in hospital settings: a scoping review. BMJ Open. 2023;13(2):e065845. https://doi.org/10.1136/bmjopen-2022-065845
24. He B, Kwan AC, Cho JH, Yuan N, Pollick C, Shiota T, et al. Blinded, randomized trial of sonographer versus AI cardiac function assessment. Nature. 2023;616(7957):520–4. https://doi.org/10.1038/s41586-023-05947-3
25. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized clinical trials of machine learning interventions in health care: a systematic review. JAMA Netw Open. 2022;5(9):e2233946. https://doi.org/10.1001/jamanetworkopen.2022.33946
26. Titano JJ, Badgeley M, Schefflein J, Pain M, Su A, Cai M, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. Nat Med. 2018;24(9):1337–41. https://doi.org/10.1038/s41591-018-0147-y
27. Nimri R, Battelino T, Laffel LM, Slover RH, Schatz D, Weinzimer SA, et al. Insulin dose optimization using an automated artificial intelligence-based decision support system in youths with type 1 diabetes. Nat Med. 2020;26(9):1380–4. https://doi.org/10.1038/s41591-020-1045-7
28. Seol HY, Shrestha P, Muth JF, Wi CI, Sohn S, Ryu E, et al. Artificial intelligence-assisted clinical decision support for childhood asthma management: a randomized clinical trial. PLoS One. 2021;16(8):e0255261. https://doi.org/10.1371/journal.pone.0255261
29. U.S. Food and Drug Administration. Policy for Device Software Functions and Mobile Medical Applications. 2022. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/policy-device-software-functions-and-mobile-medical-applications
30. Ding L, Liu C, Li Z, Wang Y. Incorporating artificial intelligence into stroke care and research. Stroke. 2020;51(12):e351–4. https://doi.org/10.1161/STROKEAHA.120.031295
31. Chen RJ, Wang JJ, Williamson DFK, Chen TY, Lipkova J, Lu MY, et al. Algorithmic fairness in artificial intelligence for medicine and healthcare. Nat Biomed Eng. 2023;7(6):719–42. https://doi.org/10.1038/s41551-023-01056-8
32. Qi M, Cahan O, Foreman MA, Gruen DM, Das AK, Bennett KP. Quantifying representativeness in randomized clinical trials using machine learning fairness metrics. JAMIA Open. 2021;4(3):ooab077. https://doi.org/10.1093/jamiaopen/ooab077
33. Giovanola B, Tiribelli S. Beyond bias and discrimination: redefining the AI ethics principle of fairness in healthcare machine-learning algorithms. AI Society. 2023;38(2):549–63. https://doi.org/10.1007/s00146-022-01455-6
34. Thompson HM, Sharma B, Bhalla S, Boley R, McCluskey C, Dligach D, et al. Bias and fairness assessment of a natural

language processing opioid misuse classifier: detection and mitigation of electronic health record data disadvantages across racial subgroups. J Am Med Inform Assoc. 2021;28(11):2393–403. https://doi.org/10.1093/jamia/ocab148

35. Yuan H, Kang L, Li Y, Fan Z. Human-in-the-loop machine learning for healthcare: current progress and future opportunities in electronic health records. Med Adv. 2024;2(3):1–5. https://doi.org/10.1002/med4.70

How to cite this article: Yuan H. Toward real-world deployment of machine learning for health care: external validation, continual monitoring, and randomized clinical trials. Health Care Sci. 2024;3:360–4. https://doi.org/10.1002/hcs2.114