Pennestrì *et al. BMC Medical Informatics and Decision Making* (2025) 25:56
https://doi.org/10.1186/s12911-025-02883-2

BMC Medical Informatics
and Decision Making

COMMENT Open Access

Sharing reliable information worldwide: healthcare strategies based on artificial intelligence need external validation. Position paper

F. Pennestrì[1*†], F. Cabitza[1,2†], N. Picerno[1] and G. Banfi[1,3]

Abstract
Training machine learning models using data from severe COVID-19 patients admitted to a central hospital, where entire wards are specifically dedicated to COVID-19, may yield predictions that differ significantly from those generated using data collected from patients admitted to a high-volume specialized hospital for orthopedic surgery, where COVID-19 is only a secondary diagnosis. This disparity arises despite the two hospitals being geographically close (within20 kilometers). While machine learning can facilitate rapid public health responses, rigorous external validation and continuous monitoring are essential to ensure reliability and safety.

Keywords Artificial intelligence, External validation, Machine learning, Orthopedics, Patient stratification, Technovigilance

The Artificial Intelligence (AI) Act was released by the European Commission on December 9, 2023 while Machine Learning (ML) methodologies gain increasing interest worldwide [1–2]. ML technologies offer to the global community of patients, clinicians and policy makers the unprecedented opportunity to collect, elaborate and share a huge quantity of data in short time. Predictive models need to be validated in different settings and contexts to be reliable. In case of sudden global health crises, such models can support infection diagnosis and screening, predict individual outcomes and adopt effective preventative measures on time, like isolating individuals in the early contagion outbreak, developing differential pathways in care access, predicting recovery curves based on patient stratification, organizing intermediate care facilities and minimize hospitalization; these measures, in turn, can save healthcare systems from collapse [3].

However, we are far from reproducing good AI systems' performance in different settings. Lack of generalizability makes the so-called reproducibility crisis a clinical (cf. poor predictions, automation bias) and moral issue (cf. iatrogenic harm), as the substantially different meaning of validation between medical literature and ML [4–6] is often neglected. In the medical context, validation generally implies confirming the consistent fulfillment of specific requirements for intended use. In contrast, ML validation tends to focus on fine-tuning model parameters or comparing model performances, also known

[†] F. Pennestrì and F. Cabitza contributed equally to this work.

*Correspondence:
F. Pennestrì
federico.pennestri@grupposandonato.it
[1] Direzione Scientifica, IRCCS Istituto Ortopedico Galeazzi, Via Cristina Belgioioso 173, 20157 Milano, MI, Italy
[2] Dipartimento di Informatica, Sistemistica e Comunicazione, Università degli Studi Milano-Bicocca, Viale Sarca 126, 20125 Milano, MI, Italy
[3] Vita-Salute San Raffaele University, Via Olgettina 58, 20132 Milano, MI, Italy

among ML practitioners and researchers as "internal validation". Internal validation is based on the same data used to develop and test a model or on data coming from the same facility. Another thing is external validation, which evaluates the model's performance under varying conditions and datasets from different settings. An externally validated model is a more robust model, considered robustness as the capability of a system (a) to maintain its level of performance under any circumstances (ISO/IEC FDIS 22989:2022), (b) to have comparable performance on inputs dissimilar to those on which it has been trained (ISO/IEC TR 24029-1). Medical validation and machine-learning validation differ significantly in both scope and execution. While machine learning models are often validated based on historical datasets and performance metrics like accuracy, external medical validation is essential to assess how these models perform in real-world and diverse clinical environments. Providing reliable and timely information across different healthcare settings is critical to address the reproducibility crisis of AI applications in healthcare.

The literature reports many cases in which robust external validation (performed across especially diverse clinical settings and patient populations) would have revealed the limitations of the models before full deployment, allowing for recalibration or model redesign. The most known cases are IBM Watson for Oncology, Deep mind Retinal Screening for Diabetic Retinopathy and Epic's Sepsis Prediction Model [7, 8]. IBM Watson for Oncology was developed to provide treatment recommendations for cancer patients. The system was initially trained on data from Memorial Sloan Kettering Cancer Center (MSKCC); however, when applied in different hospitals (particularly in Asia), it became evident that its recommendations were not always appropriate for local clinical settings [9–12]. The system had been trained and validated on a narrow dataset, which lacked diversity in terms of population, disease variants and treatment approaches. Deep Mind AI developed a deep learning model for detecting diabetic retinopathy from retinal images. Initial results published from testing in highly controlled clinical settings were promising [13, 14]. However, in real-world settings in rural Thailand, the model faced significant challenges. Factors like image quality, variability in equipment, and differences in technician expertise led to poor model performance [15–17]. Epic Systems is a widely used electronic health records company which developed a sepsis prediction model to identify patients at risk. However, research showed that its performance in real-world hospital settings was significantly poorer than expected, with high rates of false positives [18]. The model had not been thoroughly externally validated in diverse hospital settings and clinical workflows.

In all of these cases, validation accomplished across hospitals with different patient populations and clinical workflows, would have mitigated the risk of overfitting on specific hospital data and would have allowed hospitals to adapt the model for their specific treatment, diagnostic and screening protocols, ensuring that the predictions were more aligned with local clinical realities. Similarly, collecting data from severe COVID-19 patients admitted in a central hospital with entire wards specifically dedicated to COVID-19 may be different from collecting data from patients admitted to a high-volume specialistic hospital to undergo orthopedic surgery, who are diagnosed COVID-19 as a simultaneous condition, although these hospitals are geographically equivalent. Major reference centres for the treatment of common musculoskeletal conditions (e.g. joint arthroplasty in elderly patients affected by osteoarthritis or femoral head fractures), therefore, may appear highly representative of the local population in vulnerable patients and areas to COVID-19, efficiently training predictive models in short time; however, they may result clinically misleading when used on patients with different age and/or medical characteristics just 20 km far [19]. A concrete and straightforward way to ascertain external validation, therefore, is to test the system (a) on data coming from different healthcare settings (e.g., laboratories, radiological departments, hospitals), possibly from different regions and countries (cross-sectional validation); (b) on data purposely collected for the system validation at different times (ideally prospective, and ideally years apart) (longitudinal validation). If the results are replicated, then the system can be deemed externally valid and robust.

Moreover, all models– no matter how weak or robust they are– can be invalidated by concept drift (e.g., phenotype evolution) or label shift (e.g., changing the name we give to medical phenotypes, e.g. calling disease what is previously or elsewhere called symptom). To ensure that reliable clinical information is provided from different contexts, on different patients and over time, monitoring actions like techno-vigilance are needed, just like they are in case of drugs and medical devices, in order to identify, evaluate, understand and prevent the underperformance and unwanted effects of each predictive model; even more so when a certain model tested in a certain part of the world must be available to other patients in other contexts affected by the same public health emergency, as it may happen in a next pandemic.

To be more specific on the concept of techno-vigilance, we draw a parallel with well-established practices and procedures of pharmacovigilance and adapt them to the specific challenges posed by AI. Indeed, while pharmacovigilance is designed to monitor, assess and prevent adverse effects or other problems related to pharmaceuticals, similarly, techno-vigilance is proposed to monitor the safety, performance and ethical use of AI systems in healthcare. A comprehensive techno-vigilance framework for AI in healthcare begins with pre-deployment risk assessment, where clinical and technical risks are identified, and risk profiles are developed to define the AI model's reliability. Post-deployment, continuous monitoring is also essential, with mechanisms in place to track performance in real-world settings, log failures, and detect performance drifts. Oversight bodies, or techno-vigilance committees comprising multidisciplinary experts, should regularly evaluate the AI system's safety, performance and compliance with regulations. Continuous revalidation and mandatory update cycles ensure that AI models remain aligned with evolving clinical guidelines and population data, with each update undergoing rigorous external validation.

Training healthcare providers on AI model limitations, and implementing feedback mechanisms, helps identify potential issues early. Regulatory compliance is maintained through the alignment of techno-vigilance processes with standards, ensuring accountability via regulatory audits. Additionally, ethical monitoring systems are necessary to safeguard against biases, incorporating ethical oversight to ensure AI models are used responsibly and equitably. The above outlined framework parallels pharmacovigilance in drug safety, by also emphasizing continuous monitoring, regulatory alignment and ethical integrity so as to ensure that AI systems are safely and effectively integrated into healthcare. By adhering to these techno-vigilance practices, we believe healthcare systems can monitor AI models in real time, ensure their safety and efficacy across diverse populations, achieve and maintain necessarily-high ethical standards.

Surprising few articles performed external validation analyses in the last two years, although the number is increasing, considering that 192 out of 194 ML articles reporting external validation were published in the last 5 years. Assuming the Covid-19 pandemic as the last global health crisis, launching on PubMed the query ((“machine learning“[Title/Abstract]) OR (“deep learning“[Title/Abstract])) AND ((“external validation“[Title/Abstract]) OR (“externally validated“[Title/Abstract])) AND ((COVID[Title/Abstract]) OR (COVID-19[Title/Abstract]) OR (SARS-COV-2[Title/Abstract])) gave 124 results, while removing the external validation terms the result is 6,689. Most ML models reported in the literature about risk stratification, radiological discrimination and cardiological tasks [20–22] perform poorly on external data, or significantly worse than on internal data, which means they could be practically useless if not even harmful to real-world patients and practitioners [23].

Conclusions

Medical validation and machine-learning validation differ significantly in both scope and execution. While machine learning models often undergo validation based on historical datasets and performance metrics like accuracy, external medical validation is essential to assess how these models perform in real-world, diverse clinical environments. Providing reliable and timely information across different healthcare settings is critical to address the reproducibility crisis of AI applications in healthcare.

For policymakers, we advocate the establishment of clear regulatory frameworks that mandate external validation as a fundamental component of AI certification. Regulatory bodies should require AI developers to conduct multi-site external validation studies before any system can be widely deployed. Additionally, policies should encourage the creation of publicly available datasets that represent diverse populations, diseases, and healthcare environments. Such datasets would help ensure that AI models do not reinforce health disparities or perform inadequately in underrepresented communities. A systematic regulatory audit should be implemented to monitor the continuous performance of AI models over time, ensuring that models remain compliant with safety and efficacy standards as healthcare environments evolve.

For healthcare providers, the adoption of AI models must be coupled with an understanding of their limitations. Providers should actively participate in external validation processes by contributing real-world clinical data and providing feedback on model performance. Furthermore, hospitals and healthcare organizations must develop internal protocols for the ongoing monitoring of AI tools, identifying any performance deviations that could lead to adverse outcomes. Healthcare providers play a pivotal role in identifying early signals of AI model failure and should have the authority to report issues directly to regulatory bodies and AI developers, much like adverse event reporting systems in pharmacovigilance.

For AI researchers, external validation should be integrated into the model development lifecycle from the outset. Researchers must prioritize the diversity of training and validation datasets, ensuring that their models generalize across different patient populations, clinical settings, and geographies. Collaborative efforts with healthcare institutions should be encouraged to facilitate the collection of real-world data for validation purposes. AI researchers should also develop adaptive learning mechanisms within AI models, allowing them to adjust based on new validation data from different settings. Additionally, establishing a standard for transparent reporting of validation outcomes, including instances where models underperform in external settings, is essential for advancing AI trustworthiness in healthcare.

In contrast to previous global health crises, we now have the unprecedented opportunity to leverage AI models as public health measures that can stratify patient populations, activate appropriate care pathways, and organize healthcare networks in response to emergencies. The consequences of inadequate or non-existent external validation are profound, as poorly implemented models can harm patient outcomes and erode trust in AI technologies. Therefore, it is of paramount importance that certification-oriented validation studies take external validation seriously and integrate it as a critical step to ensure the safe, equitable, and effective deployment of AI systems in healthcare.

Acknowledgements
This paper was supported and funded by the Italian Ministry of Health - "Ricerca Corrente". The APC was funded by Italian Ministry of Health - "Ricerca Corrente".

Author contributions
FP, FC and GB conceptualized the paper. FP, FC and NP wrote the paper. All authors read and approved the final manuscript.

Funding
This work was supported and funded by the Italian Ministry of Health - "Ricerca Corrente". The APC was funded by Italian Ministry of Health - "Ricerca Corrente". The funding bodies played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Data availability
No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 22 March 2024 / Accepted: 20 January 2025
Published online: 04 February 2025

References

1. Dhatterwal JS, Kaswan KS, Kumar N. Telemedicine-based Development of M-Health Informatics using AI. In: Jain V, Chatterjee JM, Priyadarshini I, editors. (et al.), editors Deep learning for Healthcare decision making. New York: River; 2023. pp. 159–75.
2. Rawat V, Singh DP, Singh N, Tiwari UK. A Review of Machine Learning Techniques (MLT) in Health Informatics. In: Agrawal, R., Mitra, P., Pal, A., Sharma Gaur, M, editors International Conference on IoT, Intelligent Computing and Security. Lecture Notes in Electrical Engineering, Springer, Singapore. 2023;29.
3. Forman R, Atun R, McKee M, Mossialos E. 12 lessons learned from the management of the coronavirus pandemic. Health Policy. 2020;124(6):577–80.
4. Cabitza F, Campagner A, Soares F, 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.
5. Liu Y, Chen PC, Krause J, et al. How to read Articles that Use Machine Learning: users' guides to the Medical Literature. JAMA. 2019;322(18):1806–16.
6. Rafalo M. Cross validation methods: analysis based on diagnostics of thyroid cancer metastasis. ICT Express. 2022;8(2):183–8.
7. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ Digit Med. 2020;3:17.
8. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. Artif Intell Healthc. 2020:25–60.
9. Park T, Gu P, Kim CH, et al. Artificial intelligence in urologic oncology: the actual clinical practice results of IBM Watson for Oncology in South Korea. Prostate Int. 2023;11(4):218–21.
10. Zhou N, Zhang CT, Lv HY, Hao CX, Li TJ, Zhu JJ, Zhu H, Jiang M, Liu KW, Hou HL, Liu D, Li AQ, Zhang GQ, Tian ZB, Zhang XC. Concordance study between IBM Watson for Oncology and clinical practice for patients with Cancer in China. Oncologist. 2019;24(6):812–9.
11. Yao S, Wang R, Qian K, Zhang Y. Real world study for the concordance between IBM Watson for Oncology and clinical practice in advanced non-small cell lung cancer patients at a lung cancer center in China. Thorac Cancer. 2020;11(5):1265–70.
12. Jie Z, Zhiying Z, Li L. A meta-analysis of Watson for Oncology in clinical application. Sci Rep. 2021;11(1):5792. https://doi.org/10.1038/s41598-021-84973-5. Published 2021 Mar 11.
13. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J, Kim R, Raman R, Nelson PC, Mega JL, Webster DR. Development and validation of a deep learning algorithm for detection of Diabetic Retinopathy in Retinal Fundus photographs. JAMA. 2016;316(22):2402–10.
14. Limwattanayingyong J, Nganthavee V, Seresirikachorn K, et al. Longitudinal screening for Diabetic Retinopathy in a Nationwide Screening Program: comparing deep learning and human graders. J Diabetes Res. 2020;2020:8839376.
15. Wongchaisuwat N, Trinavarat A, Rodanant N, Thoongsuwan S, Phasukkijwatana N, Prakhunhungsit S, Preechasuk L, Wongchaisuwat P. In-Person Verification of Deep Learning Algorithm for Diabetic Retinopathy Screening using different techniques across Fundus Image devices. Transl Vis Sci Technol. 2021;10(13):17.
16. Ruamviboonsuk P, Tiwari R, Sayres R, Nganthavee V, Hemarat K, Kongprayoon A, Raman R, Levinstein B, Liu Y, Schaekermann M, Lee R, Virmani S, Widner K, Chambers J, Hersch F, Peng L, Webster DR. Real-time diabetic retinopathy screening by deep learning in a multisite national screening programme: a prospective interventional cohort study. Lancet Digit Health. 2022;4(4):e235–44.
17. Yuan A, Lee AY. Artificial intelligence deployment in diabetic retinopathy: the last step of the translation continuum. Lancet Digit Health. 2022;4(4):e208–9.
18. Wong A, Otles E, Donnelly JP, et al. External validation of a widely implemented Proprietary Sepsis Prediction Model in Hospitalized patients [published correction appears in. JAMA Intern Med. 2021;181(8):1144.
19. Campagner A, Carobene A, Cabitza F. External validation of machine learning models for COVID-19 detection based on complete blood count. Health Inf Sci Syst. 2021;9(1):37.
20. Siontis GC, Ioannidis JP et al. Response to letter by Forike: more rigorous, not less, external validation is needed. J Clin Epidemiol. 2016;69:250-1. https://doi.org/10.1016/j.jclinepi.2015.01.021. Epub 2015 Jan 31.
21. Yu AC, Mohajer B, Eng J. External validation of Deep Learning algorithms for Radiologic diagnosis: a systematic review. Radiol Artif Intell. 2022;4(3):e210064. https://doi.org/10.1148/ryai.210064.
22. Gulati G, Upshaw J, Wessler BS, et al. Generalizability of Cardiovascular Disease Clinical Prediction models: 158 Independent External validations of 104 unique models. Circ Cardiovasc Qual Outcomes. 2022;15(4):e008487. https://doi.org/10.1161/CIRCOUTCOMES.121.008487. Epub 2022 Mar 31.
23. Shah RU, Bress AP, Vickers AJ. Do Prediction Models Do More Harm Than Good? Circ Cardiovasc Qual Outcomes. 2022;15(4):e008667. https://doi.org/10.1161/CIRCOUTCOMES.122.008667. Epub 2022 Mar 31.

Publisher's note