External validation of AI models in health should be replaced with recurring local validation

Alex Youssef[1,2], Michael Pencina[3], Anshul Thakur[2], Tingting Zhu[2], David Clifton[2,4], Nigam H. Shah[5-7]

Author affiliations:
[1]Stanford Bioengineering Department, Stanford University, Stanford, CA, USA
[2]Department of Engineering Science, University of Oxford, Oxford, UK
[3]Duke University School of Medicine, Durham, NC, USA
[4]Oxford-Suzhou Centre for Advanced Research, Suzhou, China
[5]Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, CA, USA
[6]Technology and Digital Solutions, Stanford Medicine, Stanford, CA, USA
[7]Clinical Excellence Research Center, Stanford Medicine, Stanford, CA, USA

Corresponding author:
alexeyyoussef@alumni.stanford.edu

Clinical prediction models follow a standard development pipeline: model development and internal validation; external validation; and clinical impact studies. External validation studies should be followed by real-world studies evaluating the deployed models' usefulness *(1)*. However, such studies are rarely performed. External validation ends up being the de-facto test to evaluate machine learning (ML) models before deployment.

External validation is often considered the ultimate test to conclusively judge an ML model's safety, reliability, and generalizability *(2–4)*. A model that passes an external validation test on one or a few datasets is deemed generalizable, safe, and reliable. However, external validation does not guarantee generalizability or equate to model usefulness, which should be the true goal of any clinical decision-support tool. Many have demonstrated the unreliability of clinical models when tested across multiple clinical sites *(3, 5)*. Some even argued that there is no such thing as a truly validated model *(4)*. We summarize the limitations of external validation in Table 1.

Taken together, we question whether external validation should be the ultimate standard for evaluating healthcare ML algorithms. It is at odds with how ML models in healthcare are built, shared, and sold. It assumes the ability to identify target populations and validate models on representative datasets before implementation. ML solutions are brought to the market by

commercial entities seeking to implement their models across a wide range of geographies and populations. A single model externally validated on a few datasets is unlikely to deliver the desired performance across time, diverse populations, geographies, and facilities.

Data distribution shifts make this expectation of universal generalizability a particularly problematic notion in healthcare. This is especially true when model inputs are not purely biological and include operational inputs (such as those about the nature of care delivery). A model that includes operational inputs will not (and perhaps should not) generalize to all populations and healthcare facilities. In fact, if a model (such as a readmissions predictor) worked equally well across locales such as Palo Alto, Durham, and Mumbai; one has to question how that is possible given the dramatically different patient-mixes, care-protocols, and data collection processes.

After criticism about the local performance of the Epic Sepsis Model (ESM) (6), the developer announced it would fine-tune the model to each hospital's patient mix. In doing so, we move away from expecting a generalizable universal model, which the research community had argued for in the past, and implicitly embrace a site-specific localization and validation strategy.

We argue it is a fallacy to judge a model's generalizability, reliability, safety, or utility from external validation alone, especially when operational inputs are used. Using external validation to make deterministic and broad conclusions about generalizability and subsequent reliability can lead us astray. We need scalable validation techniques that work for models across healthcare facilities with vastly different operational, workflow, and demographic characteristics.

We propose that a better use of the essence of external validation would be site-specific validation performed before every local deployment and repeated on a recurring basis. Such local validation, which builds on the concept of temporal validation, would be performed (1) before deployment at a particular facility, given the novelty of the unseen local dataset, and (2) repeated over time, given the potential for performance-disruptive distribution shifts and concept drifts. This recurring local validation paradigm is new to healthcare but routine in Machine Learning Operations (MLOps), a discipline concerned with the at-scale training, deployment, monitoring, and maintenance of models. Shankar et al. (7) highlight how MLOps incorporates continuous performance monitoring and model updating (via retraining) to maintain the desired level of model performance.

Recurring local validation overcomes many shortcomings of external validation. It minimizes the Human-Computer Interface (HCI) risk, which occurs due to the heterogeneity of clinical

actions based on a fixed model recommendation. The clinical utility of a model is driven by the clinical actions taken based on model outputs. These actions and interpretations differ across different teams, facilities and over time. Only recurring local validation can account for this heterogeneity. Similarly, it can assess local usefulness outcomes such as cost-effectiveness, workflow disruptions, and fairness. We summarize how recurring local validation overcomes the limitations of external validation in Table 1. Compared to external validation, recurring local validation provides a more comprehensive and reliable evaluation of models aligned with the intent of responsible ML in healthcare.

A recurring local validation paradigm could rely on the presence of historical data to perform the initial pre-deployment tests, which can be followed by implementing a model in silent mode, where the model output is recorded and evaluated against the clinical ground truth to assess local performance (8). The model is then fine-tuned using the collected data during the silent phase. Even small amounts of local data collected during a short time frame can be valuable for localizing a model. The silent mode approach can also be adopted when historical data is unavailable, as demonstrated in a COVID-19 deterioration prediction case study (9).

Reliability across sites, time, and populations (or generalizability) is a necessary goal for healthcare ML. Aiming for universally generalizable models evaluated through external validation is unrealistic for achieving reliability. Instead, recurring local validation via MLOps is a well-traveled path to create reliable models through retraining, fine-tuning, and continual learning. Such frameworks leverage the dynamic adaptive nature of AI algorithms. Model architectures, hyperparameters, and weights can be adapted at various deployments and over time, preserving reliability while protecting performance against data shifts and concept drifts (10).

In closing, external validation is often recommended to ensure the generalizability of ML models. However, it neither guarantees generalizability nor equates to a model's clinical usefulness - the ultimate goal of any clinical decision-support tool. External validation is misaligned with current healthcare ML needs and is insufficient to establish ML models' safety or utility. Instead, we propose the MLOps-inspired paradigm of recurring local validation to maintain the validity of models and protect against performance-disruptive data variability. We should routinely and continuously perform local evaluations of models that guide care.

Conflicts of interest:
The authors declare no competing conflicts of interest.

Funding statements:
DAC is funded by an RAEng Research Chair and an NIHR Research Professorship, the NIHR Oxford Biomedical Research Centre, the InnoHK Centre for Cerebro-cardiovascular Engineering, and the Oxford Pandemic Sciences Institute. TZ was supported by the Royal Academy of Engineering under the Research Fellowship scheme.

1. N. H. Shah, A. Milstein, S. C. Bagley PhD, *JAMA.* **322**, 1351–1352 (2019).
2. A. C. Yu, B. Mohajer, J. Eng, *Radiol Artif Intell.* **4**, e210064 (2022).
3. H. Singh, V. Mhasawade, R. Chunara, *PLOS Digit Health.* **1**, e0000023 (2022).
4. B. Van Calster, E. W. Steyerberg, L. Wynants, M. van Smeden, *BMC Med.* **21**, 70 (2023).
5. C. L. Ramspek, K. J. Jager, F. W. Dekker, C. Zoccali, M. van Diepen, *Clin. Kidney J.* **14**, 49–58 (2021).
6. A. R. Habib, A. L. Lin, R. W. Grant, *JAMA Intern. Med.* **181**, 1040–1041 (2021).
7. S. Shankar, R. Garcia, J. M. Hellerstein, A. G. Parameswaran, Operationalizing Machine Learning: An Interview Study. Preprint at https://arxiv.org/abs/2209.09125 (2022).
8. A. D. Bedoya *et al.*, *J. Am. Med. Inform. Assoc.* **29**, 1631–1636 (2022).
9. A. Youssef et al., RapiD_AI: A framework for Rapidly Deployable AI for novel disease & pandemic preparedness. Preprint at https://medrxiv.org/content/10.1101/2022.08.09.22278600v2 (2022).
10. T. Granlund, V. Stirbu, T. Mikkonen, SN Comput. Sci. 2 (2021).

Table 1: The limitations of external validation and the advantages of recurring local validation

Comparison domain	External validation	Recurring local validation
Validation dataset representativeness	Datasets are often chosen based on availability rather than reflecting the populations of intended implementation	Local validation datasets represent the population of every local implementation
Dynamic nature of healthcare	External validation cannot fully capture the potential heterogeneity of data across time, geography, and facilities	Local validation is robust to the dynamic nature of healthcare because it evaluates models on every local deployment population and over time
Ability to assess clinical usefulness and fairness	External validation studies are unable to properly assess the clinical usefulness and fairness of models because those outcomes rely on the local translation of model recommendations into clinical action	Local validation can robustly assess local outcomes like clinical usefulness and fairness
Alignment with real-world machine learning implementation	A single externally validated model is unable to deliver reliable performance across varying implementation populations with significant operational differences	Local validation allows for the monitoring and localization of deployed local model instances, which ensures reliable local performance across different implementation populations
Capability to validate deep learning models	External validation does not align with the nature of deep learning models. External validation aims for universal generalizability, but deep learning models are very sensitive to data heterogeneity	Local validation allows for the localization of model instances, and so do not rely on the universal generalizability concept