Evaluation of machine learning solutions in medicine

Tony Antoniou PhD, Muhammad Mamdani PharmD MPH

■ Cite as: *CMAJ* 2021 September 13;193:E1425-9. doi: 10.1503/cmaj.210036; early-released August 30, 2021

CMAJ Podcasts: author interview at www.cmaj.ca/lookup/doi/10.1503/cmaj.210036/tab-related-content

See related articles at www.cmaj.ca/lookup/doi/10.1503/cmaj.202434 and www.cmaj.ca/lookup/doi/10.1503/cmaj.202066

Related articles have outlined problems with the development of machine-learned solutions for health care and suggested a framework for their optimal development.[1,2] The spectrum of clinical settings in which machine learning approaches have been examined for use in the health care setting has increased markedly and become more diverse in recent years. Many studies have detailed the data science and statistical bases of machine-learned tools.[2] However, comparatively few studies have focused on their evaluation and implementation.[3] We discuss how to evaluate machine-learned solutions throughout their life cycle to optimize their use and functionality in clinical practice. Internal validation — that is, ascertaining the discriminative and calibration performance of an algorithm — should be followed by evaluation of both performance and outcomes of interest in the clinical setting, as well as evaluation of the tool's implementation into existing workflows (as outlined in Figure 1).

Key points

- Evaluation of machine-learned systems is a multifaceted process that encompasses internal validation, clinical validation, clinical outcomes evaluation, implementation research and postimplementation evaluation.
- Approaches to clinical validation include comparisons of model performance with those of clinician experts and silent deployment of systems with comparisons of predictions to actual patient outcomes; clinical outcome evaluation can be done through randomized controlled trials, cohort studies, interrupted time series analyses and before-and-after studies.
- Implementation research includes qualitative and quantitative components and formative assessments and is attentive to the context in which the system is being deployed while evaluation frameworks can help teams structure their studies and analyses.
- Postimplementation evaluation is necessary to monitor for and account for threats to system performance after deployment, which may necessitate retraining and recalibration of machine-learned systems.
- A multidisciplinary team comprising data scientists, clinician experts and implementation scientists (qualitative and quantitative expertise) can help ensure that a comprehensive evaluation is undertaken before, during and after deployment.

What is the process of model or algorithm development and interval validation?

Initially, evaluation of the predictive performance of machine-learned algorithms involves assessing their discriminatory and calibration accuracy. The former quantifies the ability of the algorithm to separate individuals according to the presence or absence of a given outcome, and the latter measures how close the predicted probabilities are to actual probabilities.[4] Such experiments comprise the internal validation stage of machine-learned algorithm development and represent the majority of published reports describing machine learning in medicine.[3]

Typically, studies determining the predictive performance and accuracy of different algorithms are retrospective in nature. Large, historically labelled data sets are used to train and test algorithms.[3,5] Machine learning methods employed at this stage range from relatively familiar approaches such as linear or logistic regression to more complex neural networks and natural language processing models.[5,6] In all cases, algorithms are first "trained" on the largest portion of the data reserved for this purpose, and then evaluated on the remaining data, referred to as the test data.[3–5] When the outcome of interest is binary (e.g., disease present or absent), performance is typically reported using standard measures such as sensitivity, specificity and the area under the receiver operator characteristic curve.[5,7] For continuous outcomes (e.g., predicted dose of a medication), performance is generally quantified using measures such as the root mean squared error or mean absolute error.[8] Graphical methods, such as calibration slopes and calibration curves, can be used to assess model calibration.[9]

Although the need for clinician or stakeholder input at this technical stage of development may not be immediately apparent, clinicians can provide important insights regarding the interpretability of performance metrics and acceptable thresholds of model performance for clinical practice.[10] For example, as part of the development of a machine-learned-based early warning system predicting patient deterioration and need for intensive care within a 24-hour period, a maximum of 2 false

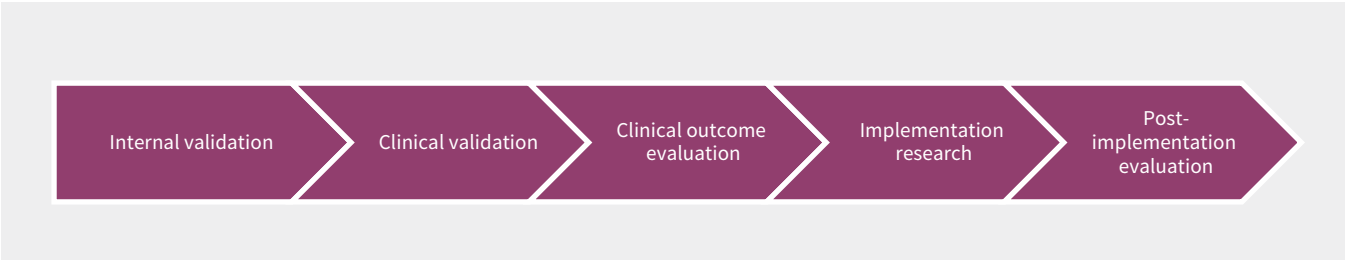

Figure 1: Evaluation life cycle of machine-learned systems in health care.

alarms per true alarm was identified by clinicians as an acceptable threshold for performance to guard against "alert fatigue."[1] Based on this requirement, it was determined that the system should have a positive predictive value of at least 0.3 while detecting as many deteriorating patients as possible.[1] Because optimal performance metrics will vary by clinical context, defining performance will therefore require consideration of clinician preferences and the care environment in which the machine-learned system will ultimately be operating.[1,10]

How should machine-learned solutions be validated clinically?

Performance of machine-learned tools on real-world data that are new to the algorithm may differ from performance during internal validation.[2] Consequently, prospective studies that compare predictions made by machine-learned algorithms with clinician predictions are required to ascertain their performance in a clinical setting. As described in our related paper, this approach was used as part of the evaluation of a machine-learned early warning system for patients on medical wards designed to identify who may require critical care; in this evaluation, we found improved sensitivity of the early warning system over prediction by clinicians.[1] Other examples include comparisons between machine-learned systems and dermatologists for diagnosing skin cancers;[11–14] diagnosis of age-related macular degeneration and diabetic retinopathy using retinal optical coherence tomography or fundus photographs;[15–17] identification of breast cancer metastases in lymph node biopsies;[18,19] and detection of polyps at colonoscopy.[20,21]

Another approach to clinical validation involves comparing the performance of a newly developed machine-learned algorithm against already validated clinical risk–scoring tools that are commonly used in clinical practice. This approach has been applied to various problems; e.g., predicting gastrointestinal bleeding and mortality after cardiac surgery.[22,23] As with approaches involving predictions by clinicians, comparisons with validated risk-scoring tools should be undertaken using data that were not part of the machine-learned model's development process.

Although many studies have shown the performance of machine-learned tools to be at least comparable to the performance of expert physicians, this is not always the case,[24] which underscores the need to conduct clinical verification studies before moving forward with more resource-intensive forms of evaluation. Clinical validation can be particularly challenging when diagnostic interrater reliability among clinicians is poor. In this context, it may be difficult to compare the discriminative performance of clinicians versus machine-learned systems, given the challenges associated with discriminating between the presence or absence of disease or associated stages of illness (e.g., remission, relapse). Potential strategies for addressing this problem include use of more concrete, measurable aspects of a specific illness (e.g., change in symptom scores or laboratory parameters) or a directly observable functional outcome (e.g., ability to return to work) rather than diagnostic labels denoting the presence or absence of disease when training models.

"Silent deployment" is another approach that may be used for clinical validation. As described in a related article, the machine-learned system runs in a silent mode and generates predictions, yet these are not communicated to clinicians and therefore do not influence care.[1] Although silent deployment typically focuses on issues related to technical deployment and workflow and does not involve clinical interventions, predictions made by the tool during silent deployment can be compared with the actual patient outcomes, which allows for estimation of the algorithm performance.

Large data sets are generally not required for the prospective validation of machine-learned algorithms. Instead, sample sizes can be estimated using established methods for studies of test accuracy.[25]

How can we establish whether machine-learned solutions improve patient outcomes?

Establishing and verifying predictive performance through internal and clinical validation studies does not answer the fundamental question of whether patients benefit from the integration of machine-learned solutions into clinical practice.[26] Generating robust evidence that supports the impact of such algorithms on patient outcomes is a prerequisite to widespread implementation in clinical practice and investment in resources and infrastructure required to continuously monitor the performance of such tools once deployed is needed.

As with other interventions, randomized controlled trials (RCTs) are the gold standard for establishing the efficacy of interventions developed through machine learning. Yet, relatively few RCTs of machine-learned interventions have been registered or published.[3,27] These include a double-blind RCT of an algorithm to detect acute neurologic events and a trial

comparing automated interpretation of cardiotocographs with usual care on clinical outcomes in mothers and infants.[28,29] Possible reasons for the dearth of RCTs in the field of machine learning include the need for large samples of patients or long durations of follow-up to show efficacy, cost and concerns regarding intervention fidelity or cross-group contamination when trials are conducted within the same institution. Although cluster RCTs could address the latter issue, these studies add to the logistical and methodological complexities inherent in multisite trials.[30,31]

Because conducting RCTs is challenging, other approaches are often used for generating evidence of clinical benefit of machine-learned systems, such as matched cohorts, quasi-experimental interrupted time series analyses, and prospective before-and-after studies.[32–34] In a related article, we described how we planned to use an observational matched cohort study design to evaluate a machine-learned early warning system in a General Internal Medicine unit, given that an RCT was estimated to require about 25 000 patients.[1] Although findings from observational studies are often considered to be a lower level of evidence than RCT findings, they provide a compromise between the needs of stakeholders and clinicians seeking timely evidence of clinical impact with machine-learned interventions and the resources required to conduct RCTs.

How can the implementation of machine-learned solutions be optimized?

Despite the potential of interventions developed using machine learning to assist with clinical decision-making and improve clinical workflow, only a few examples of successful deployment in clinical practice currently exist.[35] Moreover, studies that describe the steps taken to translate machine-learned algorithms into clinical tools are few. However, such studies are important for identifying and addressing social, ethical, organizational and logistical barriers to adoption. Implementation science — the study of methods for promoting the uptake of interventions into routine practice — should therefore be considered as fundamental as data science and clinical outcome evaluation for integration of machine-learned systems into clinical practice.[36,37] Although a detailed exposition of implementation science is beyond the scope of this article, several points merit emphasis.

In contrast to internal validation and clinical research, which emphasize the performance and efficacy or effectiveness of machine-learned solutions, implementation science research questions and outcomes focus on the process of implementation, and could include measures of intervention uptake or acceptability; they may characterize provider perceptions of the intervention on established workflows, as well as changes in processes of care.[37] In addition, understanding the context in which the machine-learned system is being implemented is important for optimizing uptake.[36] This requires addressing questions such as how to best align the system with existing workflows, how to customize the end-user interface in a manner that minimizes disruption to existing practices and which members of the care team will be interacting with the system.

Quantitative and qualitative approaches can be used for implementation research. Quantitative data can be derived through the use of structured surveys, administrative health databases, electronic health records and decision support systems, depending on the outcomes being examined.[38] Surveys can be used to ascertain facilitators and barriers to implementation, attitudes about the integration of the system in established workflows and acceptability of the intervention. Health records can be sources of information regarding intervention uptake, quality of care and costs. Qualitative methods can add depth and contextualization to quantitative approaches by examining how and why an intervention is or is not being used by clinicians, providing potential insights into interprofessional or organizational dynamics that influence uptake, and sociocultural barriers to implementation.[39] Qualitative data may be generated through in-depth interviews, focus groups, document analysis or observation, depending on the research question(s) and methodologic or theoretical orientation of the researcher.

Formative evaluations, wherein data are generated and shared with the research team and target clinicians at different stages of implementation, allow an implementation team to troubleshoot challenges arising during implementation and adapt the solution to better integrate into care processes.[40] Using an evaluation framework or theory when studying the implementation of machine-learned tools can assist researchers in structuring their studies and specifying concepts that warrant measurement. Readers are referred elsewhere for an overview of commonly used evaluation frameworks in implementation research.[41]

Why is ongoing postimplementation evaluation necessary?

Because clinical practice and processes evolve over time, the evaluation of machine-learned solutions does not end with implementation. Instead, ongoing evaluation of such systems is required to continuously monitor performance. An important threat to their performance is data-set shift, where temporal changes in clinical practice or the distribution of patient characteristics result in a data set that differs from that which was originally used to train the algorithm.[42–44] This can occur, for example, if a machine-learned algorithm is used to make clinical predictions on data from an increasingly ethnically diverse population, or a new site with a different patient population from the training data set.[2,45] Other data-related threats to system performance could include changes in the variables that were originally used in model training, such as the addition of new categories or an increasing frequency of missingness in selected variables.

Evaluating ongoing system performance may incorporate several steps,[46–49] including regularly retraining systems with the most recent data sets, comparing model performance on updated data with data currently in use and investigating discrepancies; updating outcome definitions and model inputs to align with evolving disease epidemiology, treatment or pathophysiology; generating alerts that are triggered when variable frequency distributions change; and regularly consulting with

clinical experts to monitor changes in system performance and ensure sustained clinical relevance. Where feasible, post-implementation evaluation of machine-learned solutions should be automated and scheduled at regular intervals to detect, investigate and resolve sources of system deterioration expeditiously.

Conclusion

Evaluation of machine-learned solutions is a multifaceted process that requires the expertise of data scientists, clinician experts and implementation scientists. Presently, most literature describing evaluation of these solutions remains focused on internal validation, with relatively few studies examining clinical outcomes and system implementation. This imbalance has contributed to what has been referred to as the "artificial intelligence chasm," representing the gap between the development and validation of machine-learned algorithms and their eventual use in clinical practice.[43] Additional clinical outcomes and implementation research is therefore necessary to fully realize the potential of machine learning in medicine.

References

1. Verma AA, Murray J, Grenier R, et al. Implementing machine learning in medicine. *CMAJ* 2021;193:E1351-7.
2. Cohen JP, Cao T, Viviano JD, et al. Problems in the deployment of machine-learned models in health care. *CMAJ* 2021 Aug. 30 [Epub ahead of print]. doi:10.1503/cmaj.202066.
3. Kelly CJ, Karthikesalingam A, Suleyman M, et al. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med* 2019;17:195.
4. Steyerberg EW. *Clinical prediction models: a practical approach to development, validation, and updating*. 2nd ed. Leiden (Netherlands): Springer; 2019.
5. James G, Witten D, Hastie T, et al. *An introduction to statistical learning*. New York: Springer; 2013.
6. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017;2:230-43.
7. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiology* 2003;229:3-8.
8. Chai T, Draxler RR. Root mean square error (RMSE) or mean absolute error (MAE)? – arguments against avoiding RMSE in the literature. *Geosci Model Dev* 2014;7:1247-50.
9. Van Calster B, McLernon DJ, van Smeden M, et al. Topic group "Evaluating diagnostic tests and prediction models" of the STRATOS initiative. Calibration: the Achilles heel of predictive analytics. *BMC Med* 2019;17:230.
10. Shah NH, Milstein A, Bagley SC. Making machine learning models clinically useful. *JAMA* 2019;322:1351-2.
11. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115-8.
12. Han SS, Kim MS, Lim W, et al. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. *J Invest Dermatol* 2018;138:1529-38.
13. Marchetti MA, Liopyris K, Dusza SW, et al. Computer algorithms show potential for improving dermatologists' accuracy to diagnose cutaneous melanoma: results of the International Skin Imaging Collaboration 2017. *J Am Acad Dermatol* 2020;82:622-7.
14. Haenssle HA, Fink C, Schneiderbauer R, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann Oncol* 2018;29:1836-42.
15. Kermany DS, Goldbaum M, Cai W, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell* 2018;172:1122-31.e9.
16. Burlina PM, Joshi N, Pekala M, et al. Automated grading of age-related macular degeneration from color fundus images using deep convolutional neural networks. *JAMA Ophthalmol* 2017;135:1170-6.
17. Burlina P, Pacheco KD, Joshi N, et al. Comparing humans and deep learning performance for grading AMD: a study in using universal deep features and transfer learning for automated AMD analysis. *Comput Biol Med* 2017;82:80-6.
18. Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *Am J Surg Pathol* 2018;42:1636-46.
19. Ehteshami Bejnordi B, Veta M, Johannes van Diest P, et al. Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA* 2017;318:2199-210.
20. Mori Y, Kudo SE, Misawa M, et al. Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy: a prospective study. *Ann Intern Med* 2018;169:357-66.
21. Chen PJ, Lin MC, Lai MJ, et al. Accurate classification of diminutive colorectal polyps using computer-aided analysis. *Gastroenterology* 2018;154:568-75.
22. Allyn J, Allou N, Augustin P, et al. A comparison of a machine learning model with euroscore ii in predicting mortality after elective cardiac surgery: a decision curve analysis. *PLoS One* 2017;12:e0169772.
23. Shung DL, Au B, Taylor RA, et al. Validation of a machine learning model that outperforms clinical risk scoring systems for upper gastrointestinal bleeding. *Gastroenterology* 2020;158:160-7.
24. Lehman CD, Wellman RD, Buist DS, et al.; Breast Cancer Surveillance Consortium. Diagnostic Accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015;175:1828-37.
25. Obuchowski NA. Sample size calculations in studies of test accuracy. *Stat Methods Med Res* 1998;7:371-92.
26. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *NPJ Digit Med* 2018;1:40.
27. Nagendran M, Chen Y, Lovejoy CA, et al. Artificial intelligence versus clinicians: systematic review of design, reporting standards, and claims of deep learning studies. *BMJ* 2020;368:m689.
28. Titano JJ, Badgeley M, Schefflein J, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat Med* 2018;24:1337-41.
29. Brocklehurst P, Field D, Greene K, et al. Computerised interpretation of the fetal heart rate during labour: a randomised controlled trial (INFANT). *Health Technol Assess* 2018;22:1-186.
30. Campbell MJ. Challenges of cluster randomized trials. *J Comp Eff Res* 2014;3:271-81.
31. Garrison MM, Mangione-Smith R. Cluster randomized trials for health care quality improvement research. *Acad Pediatr* 2013;13(Suppl 6):S31-7.
32. Bouaud J, Séroussi B, Antoine EC, et al. A before-after study using OncoDoc, a guideline-based decision support-system on breast cancer management: impact upon physician prescribing behaviour. *Stud Health Technol Inform* 2001;84:420-4.
33. Buising KL, Thursky KA, Black JF, et al. Improving antibiotic prescribing for adults with community acquired pneumonia: Does a computerised decision support system achieve more than academic detailing alone? — A time series analysis. *BMC Med Inform Decis Mak* 2008;8:35.
34. Harada Y, Shimizu T. Impact of a Commercial Artificial Intelligence-Driven Patient Self-Assessment Solution on Waiting Times at General Internal Medicine Outpatient Departments: Retrospective Study. *JMIR Med Inform* 2020;8:e21056.
35. He J, Baxter SL, Xu J, et al. The practical implementation of artificial intelligence technologies in medicine. *Nat Med* 2019;25:30-6.
36. Bauer MS, Damschroder L, Hagedorn H, et al. An introduction to implementation science for the non-specialist. *BMC Psychol* 2015;3:32.
37. Bauer MS, Kirchner J. Implementation science: What is it and why should I care? *Psychiatry Res* 2020;283:112376.
38. Smith JD, Hasan M. Quantitative approaches for the evaluation of implementation research studies. *Psychiatry Res* 2020;283:112521.
39. Hamilton AB, Finley EP. Qualitative methods in implementation research: an introduction. *Psychiatry Res* 2019;280:112516.

40. Elwy AR, Wasan AD, Gillman AG, et al. Using formative evaluation methods to improve clinical implementation efforts: description and an example. *Psychiatry Res* 2020;283:112532.
41. Nilsen P. Making sense of implementation theories, models and frameworks. *Implement Sci* 2015;10:53.
42. Beaulieu-Jones B, Finlayson SG, Chivers C, et al. Trends and focus of machine learning applications for health research. *JAMA Netw Open* 2019;2:e1914051.
43. Davis SE, Lasko TA, Chen G, et al. Calibration drift among regression and machine learning models for hospital mortality. *AMIA Annu Symp Proc* 2018;2017:625-34.
44. Hickey GL, Grant SW, Murphy GJ, et al. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur J Cardiothorac Surg* 2013;43:1146-52.
45. Futoma J, Simons M, Panch T, et al. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health* 2020;2:e489-e92.
46. Toll DB, Janssen KJ, Vergouwe Y, et al. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol* 2008;61:1085-94.
47. Siregar S, Nieboer D, Vergouwe Y, et al. Improved prediction by dynamic modeling: an exploratory study in the adult cardiac surgery database of the Netherlands association for cardio-thoracic surgery. *Circ Cardiovasc Qual Outcomes* 2016;9:171-81.
48. Moons KG, Kengne AP, Grobbee DE, et al. Risk prediction models: II. External validation, model updating, and impact assessment. *Heart* 2012;98:691-8.
49. Janssen KJ, Moons KG, Kalkman CJ, et al. Updating methods improved the performance of a clinical prediction model in new patients. *J Clin Epidemiol* 2008;61:76-86.

Competing interests: None declared.

This article has been peer reviewed.

Affiliations: Li Ka Shing Centre for Healthcare Analytics Research & Training (Antoniou, Mamdani), Unity Health Toronto; Li Ka Shing Knowledge Institute (Antoniou, Mamdani), Unity Health Toronto; Department of Family and Community Medicine (Antoniou), Unity Health Toronto and University of Toronto; Temerty Faculty of Medicine (Mamdani) and Leslie Dan Faculty of Pharmacy (Mamdani), University of Toronto; Institute of Health Policy, Management, and Evaluation (Mamdani), University of Toronto, Toronto, Ont.

Contributors: Both authors contributed to the conception and design of the work. Tony Antoniou drafted the manuscript. Muhammad Mamdani revised it critically for important intellectual content. Both authors gave final approval of the version to be published and agreed to be accountable for all aspects of the work.

Correspondence to: Tony Antoniou, tony.antoniou@unityhealth.to