Implementing machine learning in medicine

Amol A. Verma MD MPhil, Joshua Murray MSc, Russell Greiner PhD, Joseph Paul Cohen PhD, Kaveh G. Shojania MD MSc, Marzyeh Ghassemi PhD, Sharon E. Straus MD MSc, Chloe Pou-Prom MSc, Muhammad Mamdani PharmD MPH

■ Cite as: *CMAJ* 2021 August 30;193:E1351-7. doi: 10.1503/cmaj.202434

CMAJ Podcasts: author interview at www.cmaj.ca/lookup/doi/10.1503/cmaj.202434/tab-related-content

See related articles at www.cmaj.ca/lookup/doi/10.1503/cmaj.202066 and www.cmaj.ca/lookup/doi/10.1503/cmaj.210036

Machine learning — the process of developing systems that learn from data to recognize patterns and make accurate predictions of future events[1] — has considerable potential to transform health care. Machine-learned tools could support complex clinical decision-making and could automate many of the mundane tasks that may waste clinician time and lead to work dissatisfaction.[2] Despite growing interest in and regulatory approval of such technologies, for example smartwatch algorithms to detect atrial fibrillation,[3] to date machine-learned tools have had only limited use in routine clinical practice.[4] Developing and implementing machine-learned tools in medicine requires infrastructure and resources that can be difficult to access, such as large, real-time clinical data sets, technical skills in data science, computing power and clinical informatics infrastructure. Other barriers to adoption include challenges in ensuring data security and privacy, poorly performing mathematical models, difficulty integrating tools into existing workflows, low acceptance of machine-learned solutions by clinician users, and uncertainty about how to evaluate them.[4] In this article we outline an approach to developing and adopting machine-learned solutions in health care. Related articles discuss some of the caveats of using this technology[5] and the evaluation of machine-learned tools.[6]

Developing machine-learned solutions for clinical use requires a strong understanding of clinical care, data science and implementation science. A number of excellent frameworks support data analytics and quality-improvement initiatives, including the Cross-Industry Standard Process for Data Mining (CRISP-DM),[7] the Model for Improvement developed by the Institute for Healthcare Improvement [8] and the Knowledge to Action[9] framework. However, there is no clear, comprehensive framework specifically focused on adoption of machine-learned tools in health care. We propose a 3-phase framework to develop and implement machine-learned solutions in clinical care, illustrated by a case example (Box 1). The framework comprises an exploration phase, a solution design phase, and an implementation and evaluation phase (Figure 1). It can be used for a range of solutions, including computer vision–based projects, automation and optimization projects, and predictive analytics. The framework can also be applied when organizations are implementing machine-learned solutions that were developed elsewhere because the steps, other than model development, remain the same.

Key points

- Machine learning has the potential to transform health care, although its current application to routine clinical practice has been limited.
- Multidisciplinary partnership between technical experts and end-users, including clinicians, administrators, and patients and their families, is essential to developing and implementing machine-learned solutions in health care.
- A 3-phase framework can be used to describe the development and adoption of machine-learned solutions: an exploration phase to understand the problem being addressed and the deployment environment, a solution design phase for the development of machine-learned models and user-friendly tools, and an implementation and evaluation phase to deploy and assess the impact of the machine-learned solution.

What are the key steps of the exploration phase?

The development of successful machine-learned solutions requires a deep understanding of the problem at hand, relevant outcomes, the data that are available now and that will be available in the future, end-user needs, workflow, human factors and change management. For solutions designed to provide clinical decision support, implementation is strengthened by understanding in advance how the machine-learned solution will be paired with an evidence-based clinical intervention to improve care.

Identify the problem and build a team

The first step is to identify a problem that is important to end-users, such as clinicians or administrators, and to identify the specific, measurable outcomes they wish to change by modifying current practice. Machine-learned solutions may be geared toward replacing human effort (i.e., "do what I do"), in which case the outcomes may be time saved and measures of task performance. Alternatively, machine-learned solutions may be designed to address a clinical problem, in which case the outcome may be a measurable clinical improvement. Problems are usually first identified by end-users and then should be explored

Box 1: Case example

A failure to recognize clinical deterioration in hospital is a leading cause of unplanned patient transfer to an intensive care unit (ICU).[10] Early warning systems[11,12] can predict a patient's risk of clinical deterioration, and potentially allow clinicians to intervene earlier. Many existing early warning systems are based on traditional statistical approaches, such as logistic regression models that use a simple combination of a small number of inputs (most commonly, fewer than 10 parameters, such as vital signs), and they are prone to false-positive predictions.[13] More advanced biostatistical models may identify at-risk patients with greater accuracy.[13] However, implementation and evaluation of more advanced biostatistical or machine-learned models is uncommon.

The General Internal Medicine (GIM) inpatient service at St. Michael's Hospital, an academic health centre in Toronto, Ontario, cares for about 4000 patients each year. Roughly 7% of patients in the GIM service die or are transferred to an ICU.[14] The hospital has a well-established critical care response team, staffed by a respiratory therapist, ICU nurse and ICU physician, which can be called by ward teams to urgently assess inpatients who may require transfer to the ICU. Beginning in 2017, the hospital developed a machine-learned early warning system for the GIM service. The aim was to predict and prevent clinical deterioration to reduce mortality. Implementation and evaluation of the intervention, which was rolled out iteratively in 2020, is under way.

by a multidisciplinary team to determine whether a machine-learned solution might be appropriate. The team should include end-users who understand the clinical or operational problem and workflow; data engineers and information technology (IT) professionals who understand the available data and infrastructure and how a solution could be implemented; data scientists who understand how machine-learned models can be developed; and patients and caregivers when proposed solutions are patient-facing.

Because developing and implementing machine-learned solutions is resource intensive, great care should be taken in selecting priority projects. First, the problem should be important, which could be determined by estimating how solving the problem would improve patient health, improve patient care experience, improve provider care experience, or reduce costs. Second, a machine-learned solution must be feasible, which is determined by whether the right quantity and quality of data are available with the right timeliness, whether the problem has a reasonable chance of being modelled successfully, and whether a potential solution can be implemented within existing IT infrastructure and clinical workflow. Finally, there must be a reasonable chance of improvement associated with the interventions that will accompany the solution. Ideally, the proposed interventions are evidence based and already known to be effective. Ultimately, end-user engagement is the key to success. End-users will adopt a machine-learned solution only if it fits into their workflow and is perceived to be useful.

Understand the problem and set goals

End-users may have identified a problem that they experience regularly, but they may not understand why the problem exists or how it could be solved. The multidisciplinary team should work to understand the problem and create a theory of change, which describes their best hypothesis of how a machine-learned solution will lead to improvement. Systematic approaches to understanding clinical and operational problems have been well described, including process mapping, cause-and-effect analysis, and failure modes and effects analysis.[15] This understanding of the problem will inform the development, implementation and evaluation of the solution. As with any improvement project, the team should set clear and measurable improvement goals by defining the relevant outcomes, describing the baseline state of performance, and setting a specific target for improvement. Unique to machine-learned solutions, the team should also set performance benchmarks to define the level of model performance that would be clinically actionable and useful. It may be helpful to answer the question, "What is the current level of performance of decision-makers and by how much should it be improved for a machine-learned solution to be worthwhile?" A highly accurate model that is no better than clinical judgment will be less useful than a modestly accurate model that is substantially better than clinical judgment.

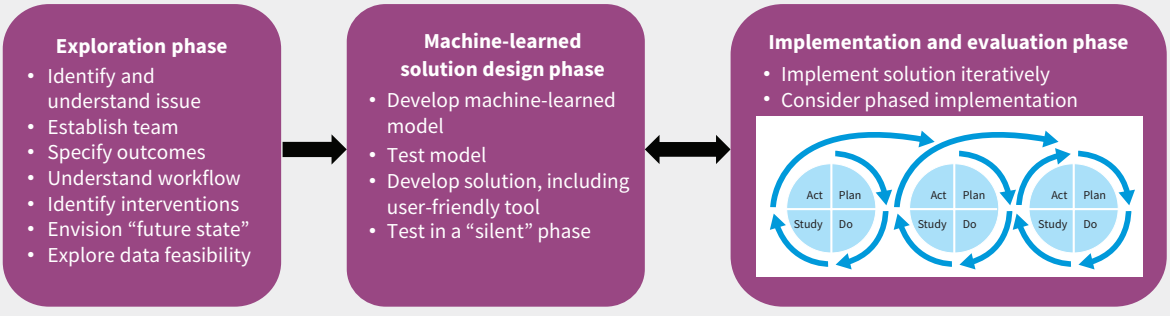

Figure 1: A framework for the development and adoption of machine- learned solutions in clinical practice.

In the case example presented in Box 1, an exploration team (Figure 2) was established to consider various clinical events that could be predicted (e.g., sepsis, acute kidney injury, readmission) to improve care for patients in the General Internal Medicine (GIM) service. Based on available data and literature review, this team created a short list of options and then consulted with the full GIM Division, hospital administrators, and 3 of the hospital's patient and family advisers before selecting clinical deterioration (i.e., death or ICU transfer) as the top priority. Data and IT experts determined that the project would be feasible. Literature review, discussions with GIM staff physicians and nurses, and a brief chart review of 10 randomly sampled[16] cases of clinical deterioration helped the team better understand the problem. The proposed theory of change was that a machine-learned early warning system might improve care by enabling earlier detection of severe illness, allowing clinicians to intervene earlier, engage in proactive conversations regarding patient preferences and goals of care, and improve the timeliness of consultation by ICU teams or palliative care teams. The team set an aim to reduce mortality in patients admitted to the GIM ward by 10% in 1 year, which was considered achievable, given other studies of early warning systems.[17]

How should machine-learned solutions be designed?

Developing a machine-learned solution involves developing and testing a machine-learned model, and then testing its initial implementation. We suggest using a framework for algorithm development and testing, such as CRISP-DM.[7] A key advantage of this approach is that it acknowledges the iterative nature of data science, which often requires cycling through 6 phases: understanding the use case, understanding the data, preparing the data, modelling, evaluating model performance, and deployment. The approach to model development is driven by several considerations, such as the problem that is being addressed; the quantity, quality and type of available data; and implementation considerations such as workflow and end-user acceptance. Developing a machine-learned solution often requires 3 complementary work streams, which could be led by 1 or more teams: model development, clinical implementation and evaluation (Figure 2). These workstreams are interrelated, as decisions made for one aspect affect the others. Focused teams can be developed for each workstream, so each receives sufficient attention and expertise, with overlapping membership to ensure coordination.

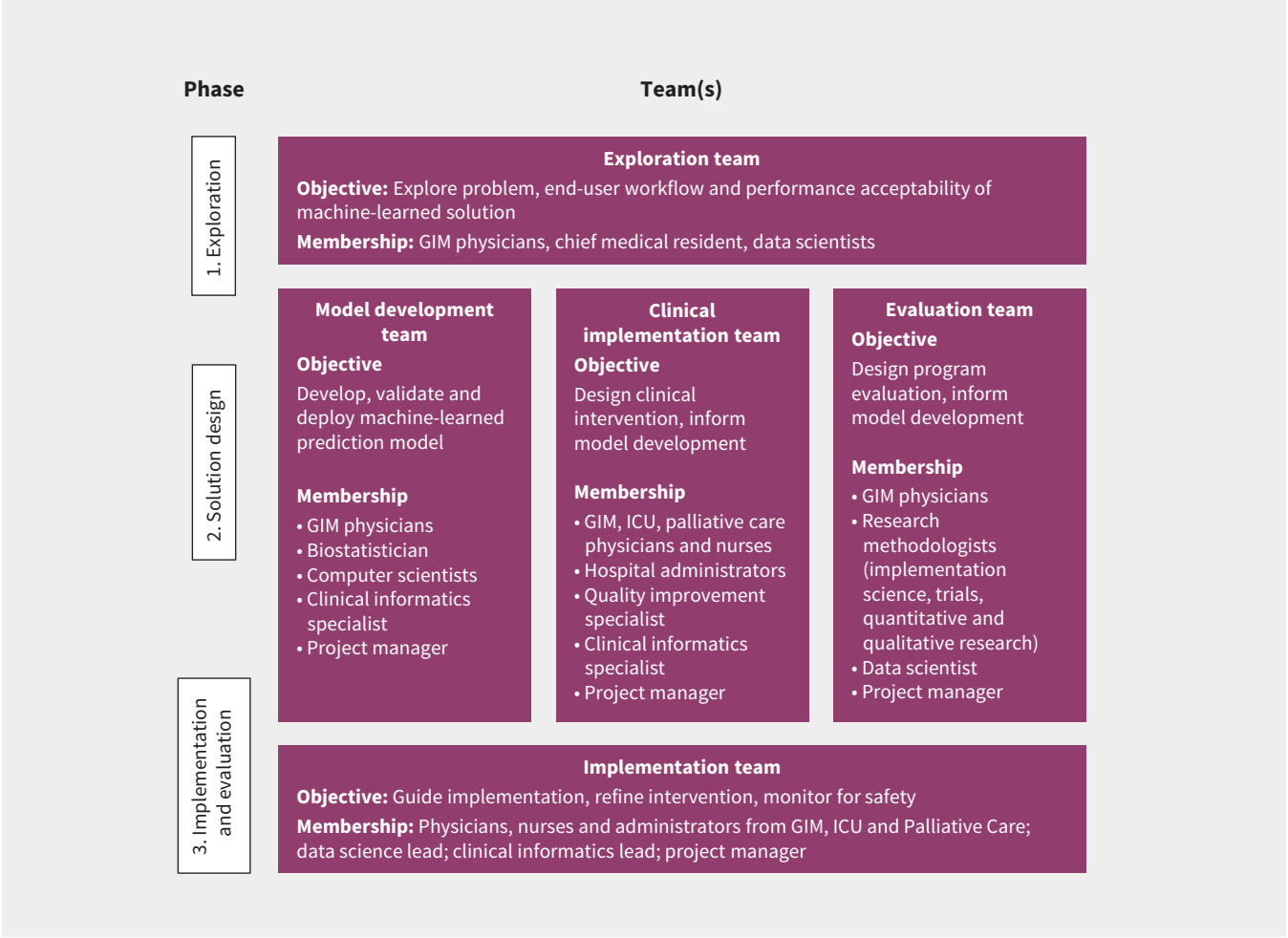

Figure 2: Team structure for each phase of development of an early warning system in the General Internal Medicine (GIM) service at St. Michael's Hospital, Toronto, Ontario. Note: ICU = intensive care unit.

Check the quality of the data

Many problems encountered when deploying a machine-learned solution can be traced back to the data used to develop the model. The quality of input data can be assessed for completeness, correctness, concordance, plausibility and currency[18] through relatively simple, automated approaches and targeted manual validation.[19] Beyond these basic data-quality metrics, it is also important to understand the outcome data that models are trained on, and whether they truly reflect the intended prediction targets. A related article discusses problems related to model training data.[5]

Design the model with implementation in mind

Data scientists have many options for developing effective models, including traditional regression techniques such as logistic regression and more modern machine-learning techniques that accommodate complex interrelationships of variables, such as neural networks.[1] Although data scientists will select a modelling approach based on the nature of the desired output and the input features,[20] the entire machine-learned solution should be designed by an interdisciplinary team with its implementation in mind.[21] In the case example (Box 1), the solution involved a prediction model, a communication system to convey patient risk to clinicians, and a clinical care pathway for high-risk patients. All aspects of the solution were designed iteratively by the 3 teams (Figure 2), with periodic input from patient and family advisers. The teams decided that the prediction model should aim for no more than 2 false alarms for every true positive alarm in order to balance the time required to assess high-risk patients with other competing demands. Thus, the data scientists set the threshold for categorizing patients as high risk at a positive predictive value of 30%, based on historical data. At this threshold, the sensitivity was 50%, which clinicians considered would be a useful proportion of cases to detect. Clinicians felt that it would be most useful to predict outcomes that were likely to occur within 24–48 hours. A much shorter window would not leave enough time to intervene, and a longer window would make it difficult for clinicians to know how to respond. Thus, the data scientists trained models to predict events in the subsequent 48 hours.

Develop a user-friendly tool

For systems designed to provide decision support, models should be incorporated into user-friendly tools that provide key pieces of useful, action-oriented information and integrate into end-user workflow. This involves collaboration between end-users and experts in process improvement, human factors, design, and change management. Engagement with end-users is critical throughout this process, although the extent of engagement will vary depending on the issue being addressed. In the case example, based on human factors principles,[22] a simple 3-level approach was selected to present actionable information to clinicians, with patients stratified into high-, medium- and low-risk groups. Clinicians receive updated patient risk predictions through the hospital's electronic signout tool and through text paging alerts. Paging alerts are sent only when patients change from lower risk levels to the highest risk level, and if a patient remains at high risk, there are no repeat alerts, thereby minimizing alarm fatigue.[23] As a result, there are typically between 0 and 2 alerts per GIM team (who usually care for 15–20 patients) per 24-hour period.

Design a clinical intervention to integrate with workflow

Introducing a new clinical tool, machine-learned or otherwise, may alter existing workflows.[24] Such changes may be planned and welcome,[25] or they may be disruptive and harmful.[26] Various strategies, including interviews, focus groups, surveys and workflow analysis, may be employed to describe existing workflows and assess barriers and facilitators to implementation of a new tool.[24,27] These can then be mapped to effective strategies to optimize implementation using approaches such as the Capability, Opportunity, Motivation, Behaviour (COM-B) model.[28] In the case example, the implementation team included clinicians and administrators with first-hand experience of the existing workflows in GIM, ICU, palliative care and clinical informatics. Additional interviews and focus groups were conducted to inform the implementation team as needed. The team considered existing resources, such as hospital protocols for escalation of care and the critical care response team when designing the intervention. The methods and timing of alerts were designed to fit within existing processes for physicians and nurses in the GIM service, ICU and palliative care. For example, model predictions are reported to charge nurses at specific times, and in a specific format, so that patient risk can be factored into nursing assignments. A clinical pathway was designed with concrete actions and time targets for physicians and nurses to respond to high-risk patients while leaving room for clinical judgment (Figure 3).

Engage end-users to establish trust

One common barrier to the adoption of machine-learned technology is whether clinicians trust the model's output.[29] One framework suggests trust can be built by demonstrating transparency, fairness and robustness of models.[30] In the case example, the team used historical data from 2011 to 2020 to develop and validate the early warning system model. Multivariate adaptive regression spline models were developed using about 100 inputs related to patient demographics, vital signs and laboratory test results; this model was chosen after experimentation with numerous modelling techniques using more than 500 input variables.[31] The large number of inputs and the complex ways they can interact make it difficult to explain the factors influencing any given prediction, although some machine-learned models may be more interpretable than others (i.e., it may be possible to report the relative importance of different predictors). It may be desirable for machine-learned models to be interpretable for some clinical applications,[32] but interpretability is not essential for establishing trust[33] and there is no consensus on the best methods to explain more complex models.[34,35] Providing detailed explanations for model predictions could even hinder clinical decision-making in some situations through information overload or creating false impressions of causality.

To establish trust in the GIM early warning system, we transparently reported to the front-line clinicians how we developed and validated these models, showing that models were not biased across patient age and sex (although there were limited sociodemographic data to explore other dimensions of fairness). We showed model robustness by validating the machine-learned models on historical cohorts using temporal split-sample validation, meaning that models trained on data from 2011 to 2019 were tested on data from 2020. We also compared model predictions to predictions made in real time by physicians and nurses about their patients over a 4-month period, to provide clinical validation of the model's potential usefulness. To encourage engagement of end-users, the initiative was championed by well-regarded senior clinical leaders, including nursing leadership and the physician heads of the GIM, ICU and Palliative Care divisions.

Engaging patients, family members and caregivers is important, particularly when developing patient-facing solutions. Engaging patients can improve the design, safety and satisfaction associated with new services.[36,37] Methods for this engagement have been well described[38,39] and should include clearly articulating the purpose of engagement, accommodating unique needs to make participation accessible, recruiting diverse partners, and embracing the opportunity for exchange between those with expert knowledge and those with lived experience. In the case example, patients and caregivers were recruited primarily from the hospital's patient and family advisory group and were consulted at various stages of the project. We chose a consultative model of engagement in order to solicit feedback on key issues, including selecting clinical deterioration as a priority, designing the clinical intervention and addressing issues related to implementation. For example, a major topic of discussion was how patients and their families should be informed about the model's predictions. These discussions led the clinical implementation team to conclude that the patient's physicians should be responsible for discussing the model's predictions when clinically appropriate and situating these in the broader context of the patient's health and treatment plan.

How should machine-learned solutions be implemented and evaluated?

Phased implementation

Widespread adoption of machine-learned solutions in health care immediately after their development is not advised. Instead, the machine-learned solution should be deployed in a "silent testing" period before formal implementation (i.e., without end-users being aware of the model predictions or recommendations). The length of this period should be determined by several factors, including the frequency of events being predicted, the nature of the specific clinical practice being targeted, and the number and

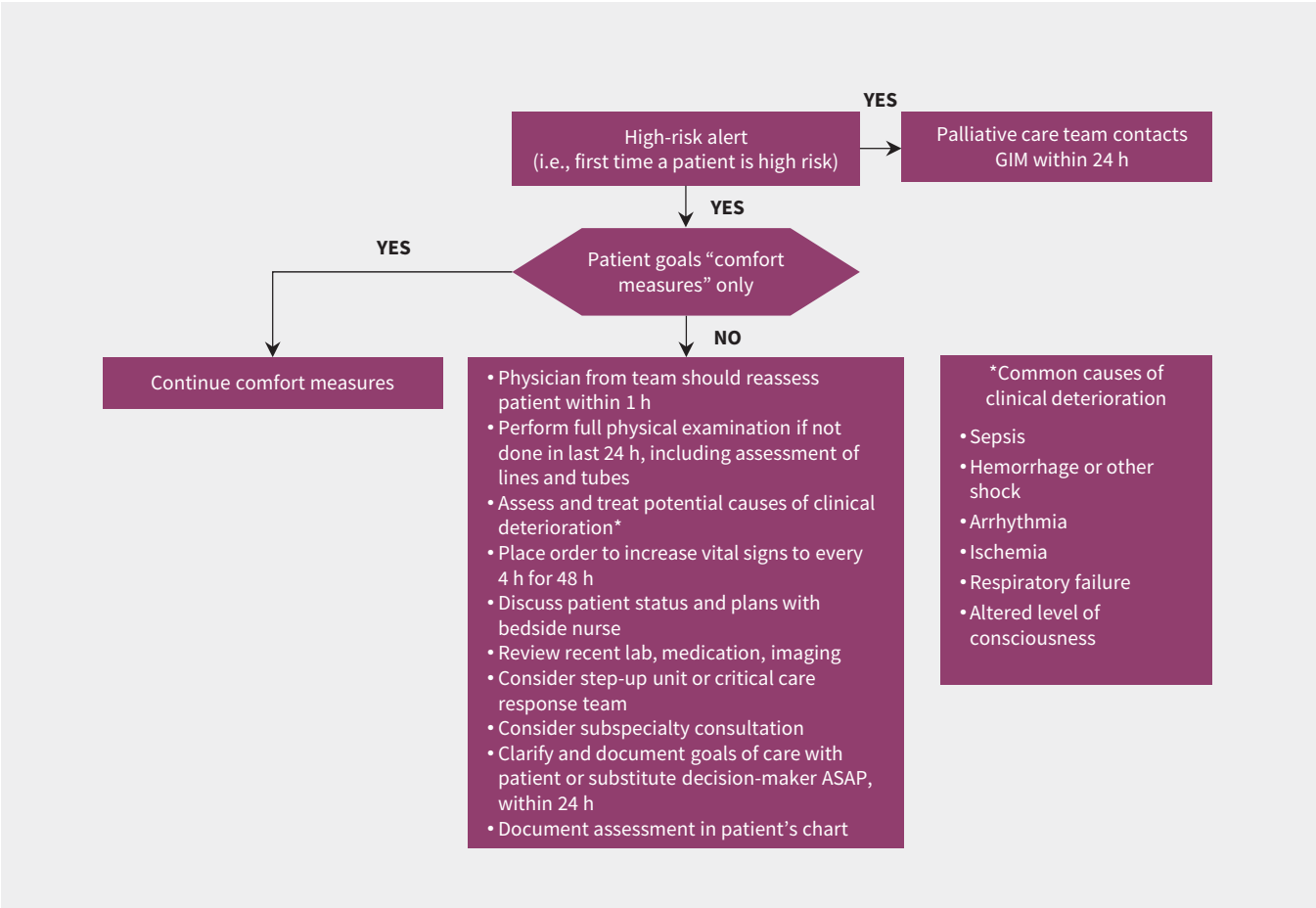

Figure 3: Clinical care pathway for patients in the General Internal Medicine (GIM) service with high predicted risk of clinical deterioration.

heterogeneity of intended end-users. This time is used to ensure that data and IT infrastructure function well and to ensure that model performance in the real-world setting is sufficient for deployment. Once successfully completed, the results of the silent trial can be reported to end-users to strengthen trust. If unsuccessful, this testing phase can prevent a potentially harmful model from being deployed, or highlight the need for refinement before deployment. In the case example, the model was silently tested in real time without communicating predictions to clinicians for 9 months. We identified and corrected several issues; for example, we corrected a computing error where the algorithm recognized "Na" (the chemical symbol for sodium) as "NA" (denoting missing values), which affected model performance.

Iterative evaluation

Given the complexity of both model development and the health care environment, we suggest applying an iterative approach using frameworks that incorporate the Plan-Do-Study-Act (PDSA) cycle,[40,41] described by the Model for Improvement developed by the Institute for Healthcare Improvement.[8] This involves "planning" the solution deployment, its aims and key measures of effectiveness and safety; "doing" the implementation on a small scale; "studying" the implementation process and impact on the stated measures; and "acting" to refine the implementation process based on the study cycle. Evaluating the implementation of machine-learned models is an iterative process — described in more detail in a related article[6] — that often requires several PDSA cycles before the solution is integrated effectively into routine workflow.

After the silent test, we launched the early warning system in the case example in a phased roll-out with 2 GIM clinical teams in August 2020, expanded to all 5 GIM clinical teams in September, and then expanded to nurses and the palliative care team in October. The phased approach allowed us to monitor and correct any unanticipated problems that might have occurred related to the machine-learned model, IT environment or clinical workflow. During implementation, the 3 project teams that led the exploration and solution design phases were collapsed into a single implementation team (Figure 2) that met weekly to review process measures and outcome measures and iteratively refine the intervention, improve adherence to the clinical pathway and address unintended consequences. We corrected issues such as erroneous alert messages, revising the alert criteria and changing the education and training processes for physicians and nurses.

Methods for evaluation

Although randomized controlled trial (RCT) designs are ideal for studying the impact of interventions, non-RCT designs such as interrupted time series methods may also be suitable. In the case example, the option of conducting an RCT was explored, but the sample size required (more than 30 000 participants would be needed to detect a 10% relative mortality reduction, given baseline mortality of about 6%) was prohibitive. A pragmatic and mixed-methods approach is being adopted instead, which includes a qualitative evaluation to identify barriers to implementation and to study the effects of the machine-learned solution on clinical practice through in-depth interviews with nurses, residents and staff physicians. Time series methods and a matched cohort design will be used to compare outcomes for patients in the intervention period to historical controls. These two approaches may help address patient-level and secular confounding, but the confounding effects of the COVID-19 pandemic will remain an important limitation. Multisite trials networks dedicated to evaluating new machine-learned technologies are needed to enable rigorous evaluation.

Conclusion

The notion that machine learning can rapidly and radically transform health care by automating mundane tasks and enhancing clinical decision-making is glamorous. Unfortunately, the reality of machine learning in health care is sobering, with many instances of poor implementations of machine-learned tools.[5] Finding machine-learned solutions that work requires careful engagement with the "messiness" of health care data and the complexity of clinical decisions and workflows. Machine learning does hold tremendous potential to meaningfully advance health care. A disciplined, inclusive, engaged and iterative approach to the development and adoption of these technologies is needed to truly benefit the patients we serve.

References

1. Liu Y, Chen PHC, Krause J, et al. How to read articles that use machine learning: users' guides to the medical literature. *JAMA* 2019;322:1806-16.
2. Sinsky C, Colligan L, Li L, et al. Allocation of physician time in ambulatory practice: a time and motion study in 4 specialties. *Ann Intern Med* 2016;165:753-60.
3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med* 2019;25:44-56.
4. Ben-Israel D, Jacobs WB, Casha S, et al. The impact of machine learning on patient care: a systematic review. *Artif Intell Med* 2020;103:101785.
5. Cohen JP, Cho 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.
6. Antoniou T, Mamdani MM. Evaluation of machine learning solutions in medicine. *CMAJ* 2021 Aug. 30 [Epub ahead of print]. doi:10.1503/cmaj.210036.
7. Chapman P, Clinton J, Kerber R, et al. CRISP-DM 1.0: a step-by-step data mining guide. Armonk (NY): SPPS; 2000. Available: https://www.the-modeling-agency.com/crisp-dm.pdf (accessed 2021 May 18).
8. How to improve. Boston: Institute for Healthcare Improvement. Available: http://www.ihi.org/resources/Pages/HowtoImprove/default.aspx (accessed 2021 May 18).
9. Graham ID, Logan J, Harrison MB, et al. Lost in knowledge translation: time for a map? *J Contin Educ Health Prof* 2006;26:13-24.
10. van Galen LS, Struik PW, Driesen BEJM, et al. Delayed recognition of deterioration of patients in general wards is mostly caused by human related monitoring failures: A root cause analysis of unplanned ICU admissions. *PLoS One* 2016;11:e0161393. doi: 10.1371/journal.pone.0161393.
11. Burch VC, Tarr G, Morroni C. Modified early warning score predicts the need for hospital admission and in-hospital mortality. *Emerg Med J* 2008;25:674-8.
12. McGinley A, Pearse RM. A national early warning score for acutely ill patients. *BMJ* 2012;345:e5310.
13. Linnen DT, Escobar GJ, Hu X, et al. Statistical modeling and aggregate-weighted scoring systems in prediction of mortality and ICU transfer: a systematic review. *J Hosp Med* 2019;14:161-9.
14. Verma AA, Guo Y, Kwan JL, et al. Patient characteristics, resource use and outcomes associated with general internal medicine hospital care: the General Medicine Inpatient Initiative (GEMINI) retrospective cohort study. *CMAJ Open* 2017;5:E842-9.

15. Quality Improvement Essentials Toolkit. Boston: Institute for Healthcare Improvement; 2021. Available: www.ihi.org/resources/Pages/Tools/Quality-Improvement-Essentials-Toolkit.aspx (accessed 2021 May 18).
16. Etchells E, Ho M, Shojania KG. Value of small sample sizes in rapid-cycle quality improvement projects. *BMJ Qual Saf* 2016;25:202-6.
17. Escobar GJ, Liu VX, Schuler A, et al. Automated identification of adults at risk for in-hospital clinical deterioration. *N Engl J Med* 2020;383:1951-60.
18. Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. *J Am Med Inform Assoc* 2013;20:144-51.
19. Verma AA, Pasricha SV, Jung HY, et al. Assessing the quality of clinical and administrative data extracted from hospitals: the General Medicine Inpatient Initiative (GEMINI) experience. *J Am Med Inform Assoc* 2021;28:578-87.
20. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018;319:1317-8.
21. Shah NH, Milstein A, Bagley Steven CP. Making machine learning models clinically useful. *JAMA* 2019;322:1351-2.
22. Phansalkar S, Edworthy J, Hellier E, et al. A review of human factors principles for the design and implementation of medication safety alerts in clinical information systems. *J Am Med Inform Assoc* 2010;17:493-501.
23. Sendelbach S, Funk M. Alarm fatigue: a patient safety concern. *AACN Adv Crit Care* 2013;24:378-86.
24. Zheng K, Ratwani RM, Adler-Milstein J. Studying workflow and workarounds in electronic health record–supported work to improve health system performance. *Ann Intern Med* 2020;172(Suppl 11):S116-22.
25. Teich JM, Merchia PR, Schmiz JL, et al. Effects of computerized physician order entry on prescribing practices. *Arch Intern Med* 2000;160:2741-7.
26. Koppel R, Metlay JP, Cohen A, et al. Role of computerized physician order entry systems in facilitating medication errors. *JAMA* 2005;293:1197-203.
27. Craig P, Dieppe P, Macintyre S, et al. *Developing and evaluating complex interventions: the new Medical Research Council guidance*. *BMJ* 2008;337:a1655.
28. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implement Sci* 2011;6:42.
29. Lee JD, See KA. Trust in automation: designing for appropriate reliance. *Hum Factors* 2004;46:50-80.
30. Asan O, Bayrak AE, Choudhury A. artificial intelligence and human trust in healthcare: focus on clinicians. *J Med Internet Res* 2020;22:e15154.
31. Nestor B, McCoy LG, Verma AA, et al. Preparing a clinical support model for silent mode in general internal medicine. Proceedings of the 5th Machine Learning for Healthcare Conference, *PMLR* 2020;126:950-72. Available: https://static1.squarespace.com/static/59d5ac1780bd5ef9c396eda6/t/5f22ccecc4a74012a31d4b4a/1596116209243/155_CameraReadySubmission_155_nestor2020preparing.pdf (accessed 2020 Sept. 14).
32. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat Biomed Eng* 2018;2:749-60.
33. Lipton ZC. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 2018;16:1-28. doi: 10.1145/3236386.3241340.
34. Payrovnaziri SN, Chen Z, Rengifo-Moreno P, et al. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *J Am Med Inform Assoc* 2020;27:1173-85.
35. Carvalho DV, Pereira EM, Cardoso JS. Machine learning interpretability: a survey on methods and metrics. *Electronics* 2019;8:832.
36. Bombard Y, Baker GR, Orlando E, et al. Engaging patients to improve quality of care: a systematic review. *Implement Sci* 2018;13:98. doi: 10.1186/s13012-018-0784-z.
37. Sharma AE, Knox M, Mleczko VL, et al. The impact of patient advisors on healthcare outcomes: a systematic review. *BMC Health Serv Res* 2017;17:693. doi: 10.1186/s12913-017-2630-4.
38. Kiran T, Tepper J, Gavin F. Working with patients to improve care. *CMAJ* 2020;192:E125-7.
39. Hamilton C, Hoens AM, Backman CL, et al. Workbook to guide the development of a patient engagement in research (PEIR) plan. Richmond (BC): Arthritis Research Canada and Vancouver: University of British Columbia; 2018. Available: http://www.arthritisresearch.ca/wp-content/uploads/2018/06/PEIR-Plan-Guide.pdf (accessed 2021 June 9).
40. Taylor MJ, McNicholas C, Nicolay C, et al. Systematic review of the application of the plan-do-study-act method to improve quality in healthcare. *BMJ Qual Saf* 2014;23:290-8.
41. Leis JA, Shojania KG. A primer on PDSA: executing plan-do-study-act cycles in practice, not just in name. *BMJ Qual Saf* 2017;26:572-7.
42. Blier N. Stories of AI failure and how to avoid similar AI fails [blog]. Amherst (MA): Lexalytics; 2020 Jan 30. Available: https://www.lexalytics.com/lexablog/stories-ai-failure-avoid-ai-fails-2020 (accessed 2021 May 18).

Competing interests: Amol Verma reports receiving a fellowship in Compassion and Artificial Intelligence from AMS Healthcare, in support of the present manuscript. Dr. Verma also reports receiving a Pathfinder Project grant from The Vector Institute (in support of the current work) and is a part-time employee of Ontario Health (which had no role in the work discussed in this paper). Amol Verma, Joshua Murray, Chloe Pou-Prom and Muhammad Mamdani led the development and implementation of the CHARTwatch early warning system at St. Michael's Hospital, Toronto, Ontario, Canada. No other competing interests were declared.

This article has been peer reviewed.

Affiliations: Unity Health Toronto (Verma, Murray, Straus, Pou-Prom, Mamdani); Li Ka Shing Knowledge Institute of St. Michael's Hospital (Verma, Straus, Pou-Prom, Mamdani); Department of Medicine (Verma, Shojania, Straus, Mamdani) and Institute of Health Policy, Management, and Evaluation (Verma, Mamdani) and Department of Statistics (Murray), University of Toronto, Toronto, Ont.; University of Alberta (Greiner); Alberta Machine Intelligence Institute (Greiner), Edmonton, Alta.; Montreal Institute for Learning Algorithms (Cohen), Montréal, Que.; Centre for Quality Improvement and Patient Safety (Shojania), University of Toronto; Sunnybrook Health Sciences Centre (Shojania); Vector Institute (Ghassemi, Mamdani) and Department of Computer Science (Ghassemi); Leslie Dan Faculty of Pharmacy (Mamdani), University of Toronto, Toronto, Ont.; Department of Radiology, Stanford University (Cohen), Stanford, Calif.

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

Correspondence to: Muhammad Mamdani, muhammad.mamdani@unityhealth.to; and Amol Verma, amol.verma@mail.utoronto.ca