Guest Editorial

Standardizing and Sharing Clinical AI Evaluation: From Local Decisions to Collective Learning

Marissa Binstock[1*], BA; Tania Tajirian[2*], MD, MHI, CCFP

[1]Digital Health Canada, Toronto, ON, Canada
[2]Centre for Addiction and Mental Health (CAMH), Toronto, ON, Canada
*all authors contributed equally

Corresponding Author:

Marissa Binstock, BA
Digital Health Canada
151 Yonge Street, Suite 1100
Toronto, ON M5C 2W7
Canada
Phone: 1 416 891 6521
Email: marissabinstock@gmail.com

Abstract

Clinical AI is increasingly being evaluated within individual health care organizations, yet much of what is learned remains local. Organizations may implement similar tools while using different measures, definitions, and approaches to evaluation, making it difficult to learn across settings. We believe local evaluation should do two things: inform immediate organizational decisions and contribute to broader learning about clinical AI. This does not require standardizing entire evaluations. A small common layer in how similar implementations are described, measured, and reported could make local findings more useful to others while preserving the context and flexibility organizations need. Drawing on our experience mapping clinical AI adoption in Canada and developing a practical evaluation toolkit, we propose a simple shift: design local evaluations not only to answer, "Is this working here?" but also, "What can others learn from what happened here?"

JMIR Med Inform 2026;14:e115607; doi: 10.2196/115607

Keywords: artificial intelligence; AI; clinical AI; evaluation; implementation; learning health systems; health care; standardization

Introduction

Clinical AI is moving quickly into routine care. Ambient documentation is probably the clearest example. In Canada, 12,019 primary care clinicians enrolled in the national AI Scribe Program between June 2025 and February 2026, and by the end of that period AI scribes were being used in more than 1.22 million encounters each month [1].

That kind of adoption creates an obvious need for evaluation. Organizations need to know whether these tools are improving workflow, saving time, being used as intended, and creating value, or if they are introducing new risks. They also need enough evidence to decide whether to continue, change, expand, or stop an implementation. But there is another question we think deserves more attention: What happens to everything organizations learn through those evaluations?

At the moment, not enough.

Organizations evaluating similar AI tools often do so independently. They choose different measures, define similar outcomes differently, describe their implementations in different ways, and may never share the results. This makes sense locally. Each organization has its own priorities, workflows, populations, and constraints. But at a system level, it means we may be missing an enormous learning opportunity. The issue is not simply that we need more evaluation of clinical AI. We also need to get more value from the evaluation that is already happening.

Local Evaluation, Collective Opportunity

We saw this problem first through our work mapping clinical AI adoption in Canada. A national environmental scan originally identified 152 publicly reported clinical AI initiatives across provinces and territories [2]. We could describe what was being implemented, where, for what

purpose, using what type of technology, and at what stage of deployment. What we generally could not determine was what happened as a result. Among the initiatives identified in the scan, publicly available information about outcomes was limited. Even where evidence was available, comparing experiences was difficult because organizations and studies measured different things or measured the same thing differently.

Ambient documentation again provides a useful example. Published evaluations have examined documentation time, after-hours electronic health record use, clinician burden and burnout, note quality, patient experience, and adoption. Reported outcomes and patterns of use differ across studies and settings [3-6]. That variation is not a problem to eliminate. It may be one of the most useful things to understand. If an AI scribe saved considerable documentation time in one setting and very little in another, we should want to know why. Was the workflow different? Was one tool integrated more effectively? Did clinicians use it more consistently? Were the patient populations or visit types different? Did implementation support differ? Right now, we are not particularly well set up to answer those questions across organizations. What if we were?

Standardize Enough to Learn From One Another

We are not suggesting that every organization conduct the same evaluation. That would be neither realistic nor particularly useful. Local context matters. The evaluation needed for an ambient documentation tool is different from that needed for a predictive model influencing clinical decisions. Organizations also differ in their priorities, capabilities, and reasons for evaluating a technology. Instead, we think there is value in exploring a much smaller proposition: standardizing the minimum, not the entire evaluation.

At its simplest, that could mean agreeing on a small common layer: a consistent way of describing what was implemented and in what context, a limited number of outcomes for similar use cases defined in the same way, and enough information about the evaluation approach to understand what the findings can and cannot tell us. For ambient documentation, a common minimum might include adoption, documentation time, clinician and patient experience, documentation quality, a safety signal such as reported incidents, and the decision made about continued use or expansion. Measures should be feasible in settings without reliable access to electronic health record audit logs. Everything beyond that could remain local.

This idea is not without precedent. Core outcome sets are used in research to establish a minimum set of outcomes that should be measured and reported while allowing researchers to add others [7]. Core measures derived from electronic health record logs have similarly been proposed, and later work has examined the consistency of these measures across studies [8,9]. Canada's national multivendor AI scribe evaluation also offers a practical example of why consistent measures and definitions matter when comparing experiences across products and settings [1].

Existing frameworks provide useful foundations. Reporting guidelines such as CONSORT-AI (Consolidated Standards of Reporting Trials–Artificial Intelligence), DECIDE-AI (Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence), and TRIPOD+AI (Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis plus Artificial Intelligence) support the reporting of clinical AI studies, including prediction model research [10-12], while implementation frameworks such as the CFIR (Consolidated Framework for Implementation Research) and RE-AIM (reach, effectiveness, adoption, implementation, and maintenance) help characterize implementation processes, context, and outcomes [13,14]. Our proposal builds on these approaches but focuses on the operational evaluations organizations undertake to inform local decisions, many of which are not published as research.

The principle is more important than the precise mechanism. If organizations evaluating similar clinical AI applications measure at least some of the same things in the same way and describe their implementations consistently, their findings become more useful to others. Standardization does not make weak evidence strong. Nor does it mean that results from different organizations can automatically be pooled. A simple before-and-after evaluation still has limitations. Context, evaluation design, implementation approach, and differences between products still matter. But without some common ground, we often cannot even get to the point of asking why results differ.

What could this change?

The most immediate benefit would be fairly simple: the next organization would know more than the last one did. Consider what we could learn as organizations evaluate ambient documentation. One organization may find that adoption falls after an initially enthusiastic pilot. Another may find that documentation time improves but after-hours work does not. Another may see different results by specialty. Another may identify an unexpected patient experience issue. Another may decide not to expand the implementation at all. Those findings are useful locally. Taken together, they could be much more useful. Patterns could begin to emerge. Organizations could see whether their experience is unusual or common. They could identify measures worth including in their own evaluations and implementation issues worth anticipating. Differences between settings could generate better questions about where particular tools create value and where they do not.

Just as importantly, we could learn more from implementations that do not succeed. Health care has understandable incentives to tell success stories, and null and negative results are underreported even in formal research [15]. An AI implementation that is stopped because adoption is poor, workflow effects are negative, or expected benefits fail to appear is much less likely to become a conference presentation or case study. Yet that experience may be

exactly what another organization needs before making the same investment. Shared evaluation could make those less visible findings part of the evidence base too. But voluntary sharing risks overrepresenting positive experiences; prospective registration of evaluations could be explored as one way to make the full range of outcomes more visible. It could also change the questions we ask about clinical AI. Instead of asking whether a particular category of AI "works," we could begin asking more useful questions: What appears to work? For whom? In which settings? Under what implementation conditions? What trade-offs are emerging? Where are results inconsistent?

Those are questions individual evaluations cannot answer particularly well on their own.

From Mapping Adoption to Understanding Impact

Our own thinking about this has evolved.

The Canadian environmental scan was intended to understand the clinical AI landscape: what was being implemented and where. The classification approach developed through the scan informed the standard taxonomy and glossary sections included in the Clinical AI Toolkit [16] of the Digital Health Canada Chief Executive Forum, which provides consistent terminology for describing clinical AI implementations.

The toolkit was developed to address a different problem. Organizations needed a practical way to evaluate a clinical AI implementation and decide what to do next. It includes a minimum viable evaluation (MVE), expanded evaluation, a key performance indicator library, implementation readiness and return on investment tools, decision guidance, and the standard taxonomy and glossary sections mentioned above [16]. The MVE helps an individual organization select a manageable set of measures for its own decision; it is not itself an agreed common outcome set. The shared minimum proposed here would require participating organizations to agree on a few comparable measures and definitions for a particular use case, potentially through a focused consensus process.

Neither the scan nor the toolkit was designed to create a national shared evidence system. But taken together, they raise an interesting possibility. The scan helps answer: What are we implementing? Local evaluation helps answer: What happened here? More standardized and shared evaluation could help us begin to answer: What are we learning across implementations? That is a different ambition, and we think it is worth exploring. It also fits naturally with the idea of a learning health system. Knowledge generated through practice should feed back into future practice [17]. Clinical AI may be particularly well suited to that approach because adoption is moving quickly, similar tools are being implemented simultaneously across many organizations, and formal research will not answer every operational question at the speed organizations need.

This is not an argument for replacing research, clinical validation, or other forms of rigorous evidence generation with operational evaluation. It is an argument against wasting what we learn through routine implementation.

Start With What Organizations Are Already Doing

There are plenty of reasons this will be difficult.

Measures that sound identical may not be calculated in the same way. Products and versions change. Electronic health record systems differ, and audit-log measures may not be available in smaller organizations, primary care, or community settings. Specialty, patient population, language, implementation support, and local capacity all affect interpretation. A common minimum must not privilege what is easiest to extract over what is clinically meaningful, including quality and safety. Smaller and rural organizations may need proportionate approaches to measurement. Differences in organizations' capacity to share findings could also bias the emerging evidence base toward larger or academic organizations with more resources.

Sharing also requires governance, not simply a common template. Small-site results may identify individual clinicians; vendor contracts can restrict disclosure; and privacy obligations, including applicable provincial health information laws, must be respected. Where Indigenous data are involved, Indigenous data sovereignty and appropriate community governance are essential. Organizations must also consider whether an evaluation constitutes quality improvement or research, as this distinction can affect ethics review requirements [18]. These issues matter. They are also a reason not to jump immediately to a large reporting or benchmarking infrastructure.

A better starting point is to learn from the evaluations already happening. Digital Health Canada has published the Clinical AI Toolkit and encourages organizations to use it and share experiences. The toolkit was developed with input from a working group that included health system and vendor representatives. Any future synthesis would need safeguards against promotional or selectively positive reporting. The invitation to share use cases is not a governed national registry. As experiences emerge, useful questions include whether organizations select similar measures, where definitions differ, which contextual details matter, and what findings can be shared responsibly.

We do not yet know the answers. We also should not assume that one shared minimum will work across every form of clinical AI. Ambient documentation may be a sensible place to start because adoption is widespread and many organizations are asking similar questions. Other applications, particularly predictive or diagnostic AI, raise different evaluation and validation requirements.

For example, a predictive model used to identify patients at risk of deterioration would call for different measures, such as calibration, false alerts, clinical response, and patient

safety, rather than documentation time. The shared principle is consistent definitions within a use case, not identical indicators across all clinical AI.

Make Local Learning Travel

Clinical AI will continue to be evaluated organization by organization. It should be. Decisions about implementation and scale need to reflect local patients, clinicians, workflows, risks, and priorities. But local does not have to mean isolated. The aim is not uniformity for its own sake. It is enough consistency to allow learning to travel. A practical next step would be to test this approach using a defined use case, such as ambient documentation. Participating organizations could agree on a small set of common measures and definitions, describe the context needed to interpret findings, and identify what can be shared responsibly. This would help establish where standardization adds value and where local variation needs to remain.

Clinical AI is being implemented too quickly, and in too many places, for every organization to learn the same lessons independently. We should continue asking, "Is this working here?" But we should also ask, "What can others learn from what happened here?" If we can make that second question part of how we approach evaluation, local assessments can become more than individual decisions. They can become building blocks for shared learning about clinical AI.

Acknowledgments

The authors thank the members of the Digital Health Canada Chief Executive Forum AI in Action Working Group, the University of Toronto Executive Master of Health Informatics (EMHI) and Master of Health Informatics (MHI) student volunteers who conducted the initial environmental scan, and the emerging professional volunteers who refined the scan dataset and contributed to the development of the Clinical AI Toolkit. Their contributions helped inform the work and perspectives discussed in this editorial.

Generative AI Use

Generative AI tools, including ChatGPT (OpenAI) and Claude (Anthropic), were used during manuscript development to support brainstorming, outlining, drafting, and language editing. The authors reviewed and revised the resulting text, verified the supporting evidence and references, and take full responsibility for the manuscript's content and conclusions.

Funding

No external funding was received for the development of the Clinical AI Evaluation Toolkit or the preparation of this editorial. MB's contributions were undertaken as part of her salaried role at Digital Health Canada.

Data Availability

Data availability is not applicable because this guest editorial does not report a new dataset or original analysis.

Authors' Contributions

Conceptualization: TT, MB
Writing – original draft: TT, MB
Writing – review and editing: TT, MB
Both authors reviewed and approved the final manuscript.

Conflicts of Interest

TT served as the volunteer champion of the Digital Health Canada Chief Executive Forum AI in Action Working Group and continues to serve as a member of the Chief Executive Forum Advisory Committee. She co-led development of the Clinical AI Toolkit discussed in this editorial. MB is employed by Digital Health Canada, which developed and published the toolkit. The authors declare no other conflicts of interest.

References

1. Canada Health Infoway releases results of national AI scribe program, demonstrating meaningful benefits for primary care clinicians. Canada Health Infoway. Jun 23, 2026. URL: https://www.newswire.ca/news-releases/canada-health-infoway-releases-results-of-national-ai-scribe-program-demonstrating-meaningful-benefits-for-primary-care-clinicians-862493354.html [Accessed 2026-11-11]
2. AI in action: transforming clinical care across Canada. Digital Health Canada. Sep 11, 2025. URL: https://digitalhealthcanada.com/ai-in-action-transforming-clinical-care-across-canada/ [Accessed 2026-10-13]
3. Stults CD, Deng S, Martinez MC, et al. Evaluation of an ambient artificial intelligence documentation platform for clinicians. JAMA Netw Open. May 1, 2025;8(5):e258614. [doi: 10.1001/jamanetworkopen.2025.8614] [Medline: 40314951]
4. Olson KD, Meeker D, Troup M, et al. Use of ambient AI scribes to reduce administrative burden and professional burnout. JAMA Netw Open. Oct 1, 2025;8(10):e2534976. [doi: 10.1001/jamanetworkopen.2025.34976] [Medline: 41037268]

5. Tan JYE, Rafi IBM, Sng GGR, et al. Impact of an ambient AI scribe among clinicians and patients: real-world prospective observational time-motion study. JMIR Med Inform. Mar 31, 2026;14:e85580. [doi: 10.2196/85580] [Medline: 41915701]
6. Preiksaitis C, Alvarez A, Winkel M, et al. Ambient artificial intelligence scribe adoption and documentation time in the emergency department. Ann Emerg Med. May 2026;87(5):569-574. [doi: 10.1016/j.annemergmed.2025.12.017] [Medline: 41665590]
7. Williamson PR, Altman DG, Bagley H, et al. The COMET Handbook: version 1.0. Trials. Jun 20, 2017;18(Suppl 3):280. [doi: 10.1186/s13063-017-1978-4] [Medline: 28681707]
8. Sinsky CA, Rule A, Cohen G, et al. Metrics for assessing physician activity using electronic health record log data. J Am Med Inform Assoc. Apr 1, 2020;27(4):639-643. [doi: 10.1093/jamia/ocz223] [Medline: 32027360]
9. Rule A, Melnick ER, Apathy NC. Using event logs to observe interactions with electronic health records: an updated scoping review shows increasing use of vendor-derived measures. J Am Med Inform Assoc. Dec 13, 2022;30(1):144-154. [doi: 10.1093/jamia/ocac177]
10. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, SPIRIT-AI and CONSORT-AI Working Group. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nat Med. Sep 2020;26(9):1364-1374. [doi: 10.1038/s41591-020-1034-x] [Medline: 32908283]
11. Vasey B, Nagendran M, Campbell B, et al. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat Med. May 2022;28(5):924-933. [doi: 10.1038/s41591-022-01772-9] [Medline: 35585198]
12. Collins GS, Moons KGM, Dhiman P, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ. Apr 16, 2024;385:e078378. [doi: 10.1136/bmj-2023-078378] [Medline: 38626948]
13. Damschroder LJ, Aron DC, Keith RE, Kirsh SR, Alexander JA, Lowery JC. Fostering implementation of health services research findings into practice: a consolidated framework for advancing implementation science. Implement Sci. Aug 7, 2009;4:50. [doi: 10.1186/1748-5908-4-50] [Medline: 19664226]
14. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health. Sep 1999;89(9):1322-1327. [doi: 10.2105/ajph.89.9.1322] [Medline: 10474547]
15. Curry S, Mercado-Lara E, Arechavala-Gomeza V, et al. Ending publication bias: a values-based approach to surface null and negative results. PLoS Biol. Sep 2025;23(9):e3003368. [doi: 10.1371/journal.pbio.3003368] [Medline: 40991668]
16. Chief Executive Forum Clinical AI Toolkit. Digital Health Canada. May 2026. URL: https://digitalhealthcanada.com/wp-content/uploads/2026/05/Digital-Health-Canada-Clinical-AI-Toolkit-v260503.pdf [Accessed 2026-10-13]
17. About learning health systems. Agency for Healthcare Research and Quality. May 2019. URL: https://www.ahrq.gov/learning-health-systems/about.html [Accessed 2026-10-13]
18. Canadian Institutes of Health Research, Natural Sciences and Engineering Research Council of Canada, and Social Sciences and Humanities Research Council. Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans – TCPS 2 (2022): Article 2.5. Government of Canada; 2022. URL: https://ethics.gc.ca/eng/documents/tcps2-2022-en.pdf [Accessed 2026-10-20]

Abbreviations

CFIR: Consolidated Framework for Implementation Research
CONSORT-AI: Consolidated Standards of Reporting Trials–Artificial Intelligence
DECIDE-AI: Developmental and Exploratory Clinical Investigations of Decision Support Systems Driven by Artificial Intelligence
MVE: minimum viable evaluation
RE-AIM: reach, effectiveness, adoption, implementation, maintenance
TRIPOD+AI: Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis plus Artificial Intelligence

Edited by Andrew Coristine; This is a non–peer-reviewed article; submitted 28.Oct.2026; initial comments to author 10.Nov.2026; final revised version received 11.Nov.2026; accepted 12.Nov.2026; published 27.Nov.2026

Please cite as:
Binstock M, Tajirian T
Standardizing and Sharing Clinical AI Evaluation: From Local Decisions to Collective Learning
JMIR Med Inform 2026;14:e115607
URL: https://medinform.jmir.org/2026/1/e115607
doi: 10.2196/115607

© Marissa Binstock, Tania Tajirian. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 27.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.