Patterns

CellPress
OPEN ACCESS

Perspective

A normative framework for artificial intelligence as a sociotechnical system in healthcare

Melissa D. McCradden,[1,2,3,7,*] Shalmali Joshi,[4] James A. Anderson,[1,5] and Alex John London[6]
[1]Department of Bioethics, The Hospital for Sick Children, Toronto, ON, Canada
[2]Genetics & Genome Biology Research Program, Peter Gilgan Center for Research & Learning, Toronto, ON, Canada
[3]Division of Clinical & Public Health, Dalla Lana School of Public Health, Toronto, ON, Canada
[4]Department of Biomedical Informatics, Department of Computer Science (Affliate), Data Science Institute, Columbia University, New York, NY, USA
[5]Institute for Health Policy, Management, and Evaluation, University of Toronto, Toronto, ON, Canada
[6]Department of Philosophy and Center for Ethics and Policy, Carnegie Mellon University, Pittsburgh, PA, USA
[7]X (formerly Twitter): @Mmccradden
*Correspondence: melissa.mccradden@sickkids.ca
https://doi.org/10.1016/j.patter.2023.100864

THE BIGGER PICTURE How to holistically evaluate artificial intelligence (AI) tools in healthcare remains a challenge. Frameworks like regulatory guidelines and institutional approaches have focused narrowly on the tool's performance alone, but we know that proper use of tools requires knowing the right way to use it and under what conditions. Tools are not neutral—they reflect our values. So, AI tools can be considered "sociotechnical systems"—meaning that their computational functioning reflects the people, processes, and environment. A "normative framework" for AI tools in healthcare supplies the practical guidance for operationalizing our values that incorporates not just the tool's properties but the systems surrounding its use to achieve benefit.

SUMMARY

Artificial intelligence (AI) tools are of great interest to healthcare organizations for their potential to improve patient care, yet their translation into clinical settings remains inconsistent. One of the reasons for this gap is that good technical performance does not inevitably result in patient benefit. We advocate for a conceptual shift wherein AI tools are seen as components of an intervention ensemble. The intervention ensemble describes the constellation of practices that, together, bring about benefit to patients or health systems. Shifting from a narrow focus on the tool itself toward the intervention ensemble prioritizes a "sociotechnical" vision for translation of AI that values all components of use that support beneficial patient outcomes. The intervention ensemble approach can be used for regulation, institutional oversight, and for AI adopters to responsibly and ethically appraise, evaluate, and use AI tools.

INTRODUCTION

It can be challenging to assess whether new artificial intelligence (AI) or machine learning (ML) healthcare applications promote the legitimate interests of patients and health systems. Stakeholders need a normative framework that can both assist them in navigating challenges and provide benchmarks against which new applications can be evaluated. Normative frameworks that provide guidance about algorithmic development tend to narrowly focus on attributes of the AI model, neglecting knowledge, practices, and procedures that are necessary to fruitfully integrate the model within the larger social systems of medical practice. The challenges and frictions associated with integrating these tools into clinical practice have been fruitfully studied by social scientists, who understand AI as part of a sociotechnical system.[1] However, this work tends to be retrospective and descriptive. We propose a normative framework for advancing the responsible integration of AI systems into healthcare that captures the status of ML models as key pieces of a larger sociotechnical system. Using the concept of an intervention ensemble, we argue that AI systems should be evaluated as one element within a larger ensemble of knowledge, practices, and procedures that are jointly necessary to ensure that these innovations advance the legitimate interests of patients and health systems.

To motivate our perspective, we turn first to the current approaches to responsible translation. Guidance surrounding the development of ML models focuses heavily on the model as the main product of translation. Work in this area includes roadmaps and frameworks for responsible translation, the regulatory landscape, institutional governance, and explainability/interpretability.[2–8] Typically, the goal in these frameworks is to identify the

scientific practices necessary to make and maintain a "good" model. Although what makes a model good takes on a number of different meanings, this work is often limited to measures that narrowly focus on characteristics of the model and its outputs. These include criteria such as reliability, reproducibility, true/false negatives/positives, and a variety of technical measures of accuracy. Although clinical relevance is often emphasized as a criterion for evaluating model quality,[9,10] this is often reduced to ensuring that the model is accurate at a prediction problem that clinicians feel is important.

Regulatory frameworks increasingly include rules requiring evidence for good clinical performance of models[11]—a welcome improvement given work demonstrating that a large proportion of FDA-approved AI systems have been granted on the basis of retrospective performance alone.[12] However, regulatory bodies have not gone so far as to specify the methodology (e.g., prospective, controlled, quasi-interventional clinical trials) by which evidence is gathered. There is substantial variability regarding how performance should be assessed, which methods of evaluation to use, and what measures of performance are best.

Other work aimed at promoting responsible AI development focuses on producing models that are trustworthy, interpretable, or explainable, which we define broadly here as any attempt to support the users' understanding of the way a model generates outputs from a set of inputs. Adjuncts like the "Model Facts" label offer information to satisfy the need for basic knowledge about the model and its validation process.[13] Efforts in post hoc explainability are sometimes proposed as answers to questions about responsible clinical decision-making,[14,15] but others have pointed out the computational[16] and ethical[17,18] limitations of current explainability methods for satisfying ethical decision-making in medicine. Others have remarked that explainability is not strictly needed to encourage clinical use of model outputs.[1] Common among trustworthy, interpretable, and explainable approaches is the presumption that understanding how a model made a particular prediction is sufficient for responsible use.

In contrast, social science work has rejected a narrow view of AI as a technical product in favor of a broader frame in which these products are part of a larger sociotechnical system.[1,13] For example, Madeleine Clare Elish's unique work documenting the social dimensions of SepsisWatch demonstrated that its efficacy was heavily reliant on the work of the nursing staff charged with encouraging care teams to attend and respond to model outputs.[19] Sandhu et al.[20] observed that a model's perceived clinical utility would be related to its overall ability to support the management of a clinical problem in a given unit rather than specific performance metrics. Henry et al.[21] remarked that clinicians often did not feel the need to understand how a model arrived at its predictions to use it effectively, but they also noted that clinicians held inaccurate beliefs about the model's capabilities. These descriptive findings demonstrate how clinicians can develop schemas to feel confident using ML tools at the point of care, predominantly drawing from their own personal observations of model performance and secondary mechanisms such as trustworthiness. However, they also highlight the lack of an evidence-based foundation for clinicians to draw from in using ML tools at the bedside. There is a limited understanding of the knowledge, practices, and procedures necessary for stakeholders to use AI systems to produce value for patients.[22]

To bridge this gap, we suggest a recognition of ML tools as one element within a larger intervention ensemble.[23–25] This broader conception connects the fields of responsible AI and social sciences. The framework we propose is normative in the sense that it is offered as a guide to both facilitate responsible development of AI systems and to evaluate the decisions and practices of stakeholders who are developing, procuring, or implementing them. We hope that this framework is also relevant to shaping emerging commitments from regulatory bodies exploring the kind of evidence that should be collected to support the oversight of AI systems. Additionally, given that recent systematic reviews identified that two-fifths of AI tools evaluated through clinical trials fail to demonstrate superiority to standard of care,[26] we hope that this framework offers a means to lend precision to prospective evaluation so as to increase the likelihood of positive results, thereby reducing research waste.[27]

THE INTERVENTION ENSEMBLE WITH HEALTHCARE ML

The intervention ensemble evolved from the recognition that interventions like drugs[23–25] and other technologies such as autonomous vehicles[28] are incapable of producing clinical benefit on their own. To realize their benefits, they must be embedded within an ensemble of knowledge, practices, and procedures that govern the use case for the intervention and include the conditions under which the intervention is likely to be effective, ineffective, or harmful and the steps that are required to monitor safety and efficacy.

Interventions, drugs, devices, and other tools cannot further the interests of patients unless they are used appropriately. For example, for pharmaceuticals, whether they have no effect, produce fatal toxicities, or confer clinical advantage is a function of a set of parameters including the indication(s) for which the drug provides benefit, the dosage at which the drug provides benefit, the window above which it is toxic and below which it is ineffective, the schedule on which the drug must be provided, and any additional diagnostic criteria needed to specify the populations most likely to benefit or be harmed by the drug. The intervention ensemble consists of the intervention itself (the drug) plus the set of parameters that modulate its effects in practice. Exploratory clinical research identifies (1) the boundaries within which the intervention is clinically useful and outside of which it is harmful and (2) approximate optimal values on key dimensions. Once these windows and optima have been identified, prospective confirmatory trials ascertain whether this intervention ensemble confers clinical benefit and under what conditions.[23–25]

Like pharmaceuticals, ML models alone do not improve outcomes for patients. Our contention is that responsible clinical development involves (1) identifying and properly characterizing the ensemble of knowledge and practices that must be understood and enacted if the system is to be used for clinical benefit and then (2) generating the evidence necessary to substantiate the claim that this ensemble is likely to produce a net benefit relative to alternative approaches in clinical practice.

To explore the idea of ML models as elements within an intervention ensemble, we conducted a narrative search of existing regulatory frameworks and reporting guidelines to identify components that each deemed important to translation of ML

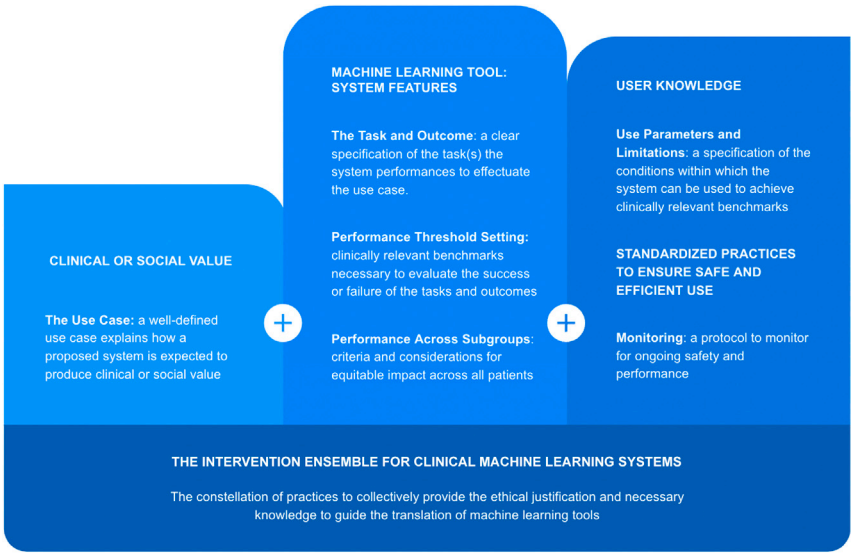

Figure 1. The intervention ensemble of clinical machine learning systems

products. While there are many proposed frameworks to facilitate translation of ML,[29] we chose to prioritize those which are predominantly oriented toward the explicit aim of evidence collection to support integration.[30–32] We also drew from published reporting guidelines evaluating AI in clinical settings[33–35] as these are explicitly geared toward establishing the required knowledge for clinical adoption. Further, drawing from some of these authors' own experiences translating models at the point of care, we considered the set of elements that clinicians felt were necessary to support them in making decisions at the bedside.

We identified the following set of elements that would surround the use of an ML model and constitute the intervention ensemble for said model's clinical use (Figure 1). Prioritizing the collection of this constellation of information as the product of a translation process for AI systems establishes the necessary and relevant set of information to guide clinical use.

(1) The use case: a well-defined use case explains how a proposed system is expected to produce clinical or social value by advancing patient interests or enhancing the capability of health systems to function more efficiently or more equitably.
(2) The task and outcome: a clear specification of the task(s) the system performs to effectuate the use case, including the way system outputs are to be integrated into clinical decision-making, practices, or procedures.
(3) Performance threshold setting: clinically relevant benchmarks necessary to evaluate the success or failure of the tasks and outcomes a system is designed to generate, their ability to achieve the intended use case, and the system's ability to produce the desired clinical benefits relative to relevant alternatives.
(4) Performance across subpopulations: criteria and considerations for equitable performance and system use across the diversity of populations whose care may be influenced by the system's outputs.
(5) Use parameters and limitations: a specification of the conditions within which the system can be used to achieve such benchmarks and outside of which its performance is expected to degrade, along with protocols or practices for implementing the system within these conditions under real-world conditions.
(6) Monitoring: a protocol for monitoring systems that are deployed in practice to ensure that their use satisfies these conditions and that changes in the clinical environment or updates to the system do not degrade its performance.

A case study approach of the IDx-DR system illustrates how an intervention ensemble might work. We chose IDx-DR because it is a widely recognized AI system that has been at the forefront of advancing health AI and has been used as a case example for many seeking to develop best practices around evaluation and oversight. Additionally, there are numerous publications to draw from surrounding the development of IDx-DR, which enriches our ability to test the applicability of the intervention ensemble. While detailed exploration of IDx-DR allows us to explore the intervention ensemble across all phases of development and implementation, we have also supplemented this case study with additional examples from the literature to enrich the discussion of the intervention ensemble. The selection of IDx-DR should not be taken as an endorsement specifically, nor are we suggesting it sets the standard.

IDx-DR is an FDA-approved AI system intended to screen for and detect mild diabetic retinopathy.[36] The outputs are provided at the point of care accompanied by a recommendation for further evaluation by a specialist or a 6-month follow-up scan. IDx-DR's approval followed a large observational trial testing the accuracy of its diagnostic properties.[36] Table 1 depicts an intervention ensemble for IDx-DR.

A well-defined use case

Per the FDA, "IDx-DR is indicated for use by health care providers to automatically detect more than mild diabetic retinopathy (mtmDR) in adults diagnosed with diabetes who have not been previously diagnosed with diabetic retinopathy." The use case thereby clearly defines a particular population (adults diagnosed with diabetes who have not been previously diagnosed with DR) and the task the system performs (detect pathologies of the eye that are established as valid indicators of mtmDR). Outside of these parameters (e.g., used to detect DR in patients without diabetes or to detect other eye pathologies), the use of IDx-DR would not be indicated based on the same evidence.

Ensuring a tight, logical link between the computational task and the use case is ideal. Passi and Barocas[37] describe "problem formulation" wherein the knowledge about the label being

predicted should be legitimately linked to the clinical problem to be addressed. Defining the use case can be relatively straightforward or it can be more complicated. Though there are "gold standards" for a plethora of disorders, there are few tests that are perfectly sensitive and specific.[38] In other contexts, such as predicting short-term mortality risk or long-term benefit, there may not be an objective expert consensus on how to define the outcomes. The use case as defined sets out the precise scope of the ML tool's application.

A clear specification of the relationship between the use case and the desired clinical benefit

The task performed by an ML model must be integrated into a larger chain of actions and decisions in a way that plausibly generates benefit. Because the relationship between diagnosis in a primary care environment and a relevant patient outcome (e.g., disease progression, time to access specialist care) was not directly tested in the trial of IDx-DR, an observational trial allows us only to hypothesize about the potential benefit. Keane and Topol[39] note that while observational studies are valuable, "such studies will not address the issue of clinical effectiveness—do patients directly benefit from the use of such AI systems?"

There are many examples of diagnostic aids, tools, and systems that demonstrate strong accuracy but have failed to yield benefits to patients. Advanced screening for various cancers through the use of computer-aided detection is emblematic of this gap: we can reliably identify abnormalities, but given how many are benign, identification itself might only result in increased anxiety, low-value testing, and waste of healthcare dollars rather than a benefit to patients.[40–42] Accordingly, to ensure appropriate use of healthcare resources and to practice evidence-based adoption of novel technologies, many strongly advocate for testing AI's systems through prospective, interventional trials prior to scaled adoption.[39,43–45]

Prospective interventional trials are increasingly pursued to test the association between the use of AI systems and a relevant clinical outcome. These can be patient-centered (e.g., mortality) or clinician/workflow-centered. As an example of the latter, BoneXpert is a system that automates bone age calculation superior to the current standard and significantly speeds up a radiologist's workflow.[46,47] On the patient-centered side, recent studies have sought to explore the relationship between the use of these tools and in-hospital mortality.[48] Ensuring clear articulation of what evidence is gathered during a particular clinical study, and how this evidence contributes to knowledge of the reliable conditions of the system's use, is central to responsible use of AI.

Clinically relevant benchmarks necessary to evaluate the success or failure of the use case and its ability to produce the desired clinical benefit

IDx-DR's outputs were defined to be consistent with established, consensus-based grading protocols.[36] The confirmation of the outputs was then established against expert performance by having images interpreted by three expert readers masked to the AI output, where a majority voting paradigm established the final diagnosis. The pre-set thresholds to define the success or failure of IDx-DR's diagnostic capabilities accounted for poor image quality and defined success as exceeding 75% for sensitivity and 77.5% specificity with an appropriate sample size. The authors contextualize this performance in light of reports that under similar parameters, board-certified ophthalmologists achieve sensitivity rates between 33% and 73%.[36]

Similarly, for BoneXpert, clinical evaluations first compare the accuracy of its bone age estimation against the performance of radiologists using the gold standard approach.[47,49] Secondly, they assessed the amount of time radiologists spend performing the task manually compared with the use of the tool. Both time to task completion and accuracy of the task are the benchmarks by which the system is judged to succeed or fail.

Criteria for equitable clinical performance across the diversity of populations on which the system is likely to be used

While it is important to ensure that accuracy is established against a reference standard or current practice, it is nonetheless important to question whether the status quo itself is effective for all patients. Health disparities are noted in diagnosis, prognosis, and access patterns within medicine. When AI systems replicate these patterns accurately, at scale, we risk further entrenching disparities.[50,51] It is therefore essential to evaluate the success or failure of an ML tool within the entire population in which it will be used, both in aggregate and in relevant subgroups.[50,52,53] Prospective clinical evaluation focused on patient outcomes can more reliably identify whether a model's impact is equitable or not. For example, in some cases, algorithmic approaches might be preferable to an existing standard.[54] However, confirmation on patient outcomes is needed to assess whether a "fair" algorithm translates to fair treatment.[53]

The trial report for IDx-DR identifies no significant effects in the performance of the system observed according to race, ethnicity, or sex.[36] They note a mildly increased specificity among those over 65 years of age. The authors described the demographic range of participants in the study according to age, sex/gender, and ethnicity/race and stated that the biomarkers of DR are considered "racially invariant." Notably, it must be acknowledged that these are somewhat imprecise terms that are proxies for factors that causally influence the outcomes of interest. For example, "sex" is often captured by either the gender or sex specified on a health card or insurance documentation or by the clinician's impression. These are distinct from features that may well influence outcomes, such as hormones, experiences of sexism, anatomy, etc., though reducing performance to certain markers risks inappropriately essentializing differences as a function of patient identity.[55] Considering what is directly measured with a given label forms a part of a holistic assessment of the overall fairness properties of a given system.[52]

Health equity scholars are re-asserting their long-standing advocacy for moving away from a neutral approach that fails to recognize differences between patients on the basis of demographic factors.[56–58] The importance of disaggregating prospective (clinical) model performance according to patient groups is increasingly recognized in medicine.[50,59] Clinical trial reporting guidelines include provisions for dis-aggregated reporting as a standard item.[33] In one example, while BoneXpert performs well overall, a prospective study noted a higher error rate among girls

Table 1. The intervention ensemble for IDx-DR: Linking the intended goals and benefits of the system with the evidence base to warrant empirical claims

	Rationale for thinking this system can be incorporated into practices and procedures that provide a specific benefit to patients/health systems	Evidence base to warrant relevant claims
Use case	detection of mtmDR in non-expert settings can facilitate timely referral for specialist-level care in the hopes of obtaining early treatment and minimizing diabetes-related vision complications	real-world impact on aggregate outcomes of health system efficiency (e.g., access to specialist care, speed of referral) has not yet been studied; no patient outcomes have yet been directly evaluated, such as access to DR-related care or a reduction in diabetes-related ophthalmological problems
Task and outcomes	automated detection of mtmDR among adults with diabetes not previously diagnosed with DR; system outputs are (1) positive for mtmDR, recommended referral to ophthalmology, or (2) negative for mtmDR, recommend re-screening in 12 months	prospective evaluation of the true clinical accuracy of the system's ability to detect mtmDR across 10 primary care sites; the outcome assessed was the system's performance against the reference standard, established by an expert panel; established the ability of non-specialist personnel to use IDx-DR in a clinical setting to detect mtmDR
Performance threshold setting	comparison to ophthalmologist accuracy rates at detecting mmDR under analogous conditions, as established by previous studies	evaluation demonstrated the accuracy of mmDR identification to be as follows: sensitivity 87%, specificity 90%, positive predictive value 73%, negative predictive value 96% (based on a prevalence of 24% for minimal DR)
Performance across subgroups	evaluate system performance across a range of relevant patient characteristics to detect differences in performance in anticipated subgroups	subgroup analysis for sensitivity and specificity by race, ethnicity, and sex to be equivalent; mild increase in specificity for adults >65 years of age; other metrics (e.g., failure case analysis) not reported
Use parameters and limitations	evaluate the relevant population and establish pertinent contraindications, warnings, standard operating procedure requirements, quality control measures, and conditions where the system fails	trial included patients with diabetes not previously diagnosed with DR, using standardized camera and equipment, using non-specialist technicians from study sites with 4 h of training, in a primary care setting; evaluated rate of images judged to be of insufficient quality and number of attempts to capture a sufficient image
Monitoring	practices necessary to monitor system performance, anticipate distribution shift, or assess performance after system updates	described as a "locked algorithm"; we were unable to find documentation in the literature regarding ongoing safety assessments or whether the system is updated over time with new data

IDx-DR, product name for the AI tool; mtmDR, more-than-mild diabetic retinopathy.

living in India.[60] As a non-demographic example, a hip fracture algorithm was noted to perform less accurately when the bone or joint in question had some abnormality.[61] Such granular data collection can more precisely inform the intervention ensemble.

A specification of the conditions within which the system can be used to achieve such benchmarks and outside of which its performance is expected to degrade

An important component of social value is that the ML model's use is valuable at a specific point in the care pathway. For example, Oakden-Rayner et al. identified that a model intended to detect the presence of respiratory conditions performed considerably better on patients with a chest tube—a treatment for respiratory conditions.[62] The implication is that a deployed model would be less accurate earlier on in the clinical pathway (where the identification of respiratory issues may be more likely to result in a net benefit) and more accurate once treatment has already commenced for the conditions for which the model is purportedly being used to identify.

For IDx-DR, the trial report contains very clear information about the conditions under which the system was evaluated.[36] Adult patients with asymptomatic, diagnosed diabetes were

evaluated in a primary care clinic by non-experts who received a standardized level of training to perform the screen. The conditions surrounding the image capture are also standardized, including the camera, patient positioning, etc. This sort of standardization and transparent reporting may foster better understanding of the generalizability of models and model approaches.[63,64]

A protocol for monitoring systems that are deployed in practice to ensure that their use satisfies these conditions and that updates to the system improve and do not degrade its performance

Much like prospective postmarketing surveillance is needed in the context of pharmaceuticals to assess ongoing safety, effectiveness, and adverse event monitoring, AI systems require analogous postdeployment monitoring for safety.[65,66] Algorithmovigilance is particularly important for AI due to its sensitivity to local contexts, susceptibility to data shifts, changes in patient-level patterns, and the effects due to changes in environmental factors (e.g., policy changes, seasonality, etc).[67] Practice shift can occur where AI systems may be used in patient populations for whom they were not initially intended, causing adverse events.[65] Notably, algorithms will differ in the extent to which they are susceptible to drifts. Since IDx-DR is a ''locked'' algorithm, for example, we are unaware of documented discussions of such protocols in this context.

Model performance can deteriorate for a variety of reasons. ''Distribution shift'' occurs when the distribution of features obtained during the training and testing of a model shift or change in such a way that the model no longer performs as expected. The cause of distribution shifts are myriad, including changes in the patient population; changes in data acquisition devices; software updates to data storage systems like electronic health records (EHRs); seasonality of diseases; clinician and patient incentives; practice recommendations; and adverse events like the COVID-19 pandemic.[68] ML systems can also suffer from runaway feedback loops. When data collected based on model predictions are used to update ML models, model decisions can be significantly biased, as identified in the case of predictive policing, where the model repeatedly recommends policing in Black neighborhoods irrespective of actual crime rate.[69] Less is understood about the implications of such feedback loops in the healthcare context.[70] Due to the heterogeneity and shifts in patient populations, maintaining institutional governance will remain important.[8] The intervention ensemble for a given system is not static and should be updated as relevant.

DISCUSSION

Treating ML tools as one component of a larger ensemble of knowledge, practices, and procedures that make up a useful medical intervention broadens the scope of evaluation for clinical translation. The approach helps researchers and clinicians guide the design and conduct of prospective evaluation to provide credible evidence of clinical utility.[27] The clear pre-specification and analyses of the various parameters surrounding a model's use can improve the trustworthiness of the research process while minimizing research waste and increasing the likelihood of a positive trial result.[27] Additionally, the intervention ensemble (rather than the model alone) can constitute a more patient-centered unit for monitoring, education, and governance.[8]

A great deal of work has revolved around demonstrating accuracy as a necessary and sufficient condition for clinical use. Once accuracy is established, researchers turn to building the trust and acceptance of the user (via, for example, explainable interfaces). But accuracy alone does not establish clinical effectiveness, and the ''likeability'' of a system is not a morally significant metric of trustworthiness. The intervention ensemble concept bridges the gap between the operating characteristics of a model and the relevant information clinicians need to advance the interests of patients. Clinicians do not require tools to be perfect in order to use them; rather, they need to know when they work well, when they do not, and how their net clinical advantage compares to other clinical alternatives.

Notably, the intervention ensemble defended here does not require AI systems to be interpretable by design or explainable by another system. As argued previously,[18] many interventions in medicine lack these properties in the sense that their clinical benefits have been demonstrated in well-designed trials but we do not know the precise mechanism by which they bring about that effect. We aim to provide standards for AI systems that are symmetrical with those applied to other medical interventions—stakeholders require the knowledge necessary to implement them in practice, and they require credible evidence that in doing so, they will advance patient interests relative to available alternatives. This is not to say we oppose interpretable systems—like others, we regard transparency and interpretability as desirable features of AI systems. However, we remain concerned that discourse surrounding AI sometimes overstates the value of interpretability and explainability, as though these conditions are necessary or sufficient for responsible deployment of AI systems. Our position is that, although these traits are desirable, they are neither necessary nor sufficient *in and of themselves* for deployment, at least insofar as these traits require knowledge of how algorithms work that goes beyond what has been described in the present framework.

Finally, we have developed this framework with the hope that it will assist in the responsible development and implementation of clinical AI. At this time, however, we have not yet evaluated its impact (e.g., can use of the intervention ensemble prevent over- and under-reliance on ML outputs, promote ML acceptance, or offer acceptable transparency to patients?). This is a task for future work. Future work might also explore its generalizability across contexts or its inclusivity to emerging variations of AI (e.g., generative models). We are interested in feedback from our colleagues on the utility of this framing.

Conclusion

Scholars have noted the chasm between current practices for developing ML systems in medicine and the evidentiary needs of key stakeholders. Bridging this gap is necessary to ensure that research in this area generates the benefits necessary to improve patient outcomes, improve the delivery of health services, reduce unwarranted variation in practice, and ensure that practices are grounded in credible evidence of safety and efficacy.[71] The intervention ensemble concept can improve the validation of AI systems by better meeting the information needs of key stakeholders, from clinician users to patients and health

systems administrators. Understanding ML models as one component of a larger intervention ensemble encourages stakeholders to specify the use case, its relationship to a clinical benefit, the performance threshold setting, criteria for equitable performance across subpopulations, parameters and limitations, and a protocol for monitoring. These components together can provide the necessary foundation for beneficial use and holistic governance of ML systems.[8]

ACKNOWLEDGMENTS

We are grateful to the anonymous reviewers whose comments helpfully enriched the discussion in this manuscript.

AUTHOR CONTRIBUTIONS

All authors contributed intellectually to the development of ideas, analysis, and drafting the manuscript.

DECLARATION OF INTERESTS

M.D.M. receives research funding support from Canadian Institutes of Health Research (CIHR), the Edwin Leong Centre for Healthy Children, the Dalla Lana School of Public Health, and the SickKids Foundation.

REFERENCES

1. Sendak, M., Elish, M.C., Gao, M., Futoma, J., Ratliff, W., Nichols, M., Bedoya, A., Balu, S., and O'Brien, C. (2020). 'The human body is a black box': supporting clinical decision-making with deep learning. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Association for Computing Machinery), pp. 99–109.

2. Chen, P.-H.C., Liu, Y., and Peng, L. (2019). How to develop machine learning models for healthcare. Nat. Mater. *18*, 410–414. https://doi.org/10.1038/s41563-019-0345-0.

3. Sendak, M.P., D'Arcy, J., Kashyap, S., Gao, M., Nichols, M., Corey, K., Ratliff, W., and Balu, S. (2020). A path for translation of machine learning products into healthcare delivery. Euro. Med. J. Innov. *10*, 19-00172. https://doi.org/10.33590/emjinnov/19-00172.

4. Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V.X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., et al. (2019). Do no harm: a roadmap for responsible machine learning for health care. Nat. Med. *25*, 1337–1340.

5. He, J., Baxter, S.L., Xu, J., Xu, J., Zhou, X., and Zhang, K. (2019). The practical implementation of artificial intelligence technologies in medicine. Nat. Med. *25*, 30–36. https://doi.org/10.1038/s41591-018-0307-0.

6. van de Sande, D., Van Genderen, M.E., Smit, J.M., Huiskens, J., Visser, J.J., Veen, R.E.R., van Unen, E., Ba, O.H., Gommers, D., and Bommel, J.v. (2022). Developing, implementing and governing artificial intelligence in medicine: a step-by-step approach to prevent an artificial intelligence winter. BMJ Health Care Inform. *29*, e100495.

7. McCradden, M.D., Anderson, J.A., A. Stephenson, E., Drysdale, E., Erdman, L., Goldenberg, A., and Zlotnik Shaul, R. (2022). A Research Ethics Framework for the Clinical Translation of Healthcare Machine Learning. Am. J. Bioeth. *22*, 8–22.

8. Kim, J.Y., Anderson, J.A., A. Stephenson, E., Drysdale, E., Erdman, L., Goldenberg, A., and Zlotnik Shaul, R. (2023). Organizational governance of emerging technologies: AI adoption in healthcare. In 2023 ACM Conference on Fairness, Accountability, and Transparency (ACM). https://doi.org/10.1145/3593013.3594089.

9. Kelly, C.J., Karthikesalingam, A., Suleyman, M., Corrado, G., and King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. BMC Med. *17*, 195.

10. Lindsell, C.J., Stead, W.W., and Johnson, K.B. (2020). Action-Informed Artificial Intelligence-Matching the Algorithm to the Problem. JAMA *323*, 2141–2142.

11. Unsworth, H., Dillon, B., Collinson, L., Powell, H., Salmon, M., Oladapo, T., Ayiku, L., Shield, G., Holden, J., Patel, N., et al. (2021). The NICE Evidence Standards Framework for digital health and care technologies - Developing and maintaining an innovative evidence framework with global impact. Digit. Health *7*, 20552076211018617.

12. Wu, E., Wu, K., Daneshjou, R., Ouyang, D., Ho, D.E., and Zou, J. (2021). How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. Nat. Med. *27*, 582–584.

13. Sendak, M.P., Gao, M., Brajer, N., and Balu, S. (2020). Presenting machine learning model information to clinical end users with model facts labels. NPJ Digit. Med. *3*, 41.

14. Amann, J., Blasimme, A., Vayena, E., Frey, D., and Madai, V.I.; Precise4Q consortium (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. BMC Med. Inf. Decis. Making *20*, 310.

15. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., et al. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. Minds Mach. *28*, 689–707. https://doi.org/10.1007/s11023-018-9482-5.

16. Ghassemi, M., Oakden-Rayner, L., and Beam, A.L. (2021). The false hope of current approaches to explainable artificial intelligence in health care. Lancet. Digit. Health *3*, e745–e750.

17. McCradden, M.D. (2021). When is accuracy off-target? Transl. Psychiatry *11*, 369.

18. London, A.J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Cent. Rep. *49*, 15–21.

19. Elish, M.C. (2018). The stakes of uncertainty: developing and integrating machine learning in clinical care. Ethnographic Praxis *2018*, 364–380.

20. Sandhu, S., Lin, A.L., Brajer, N., Sperling, J., Ratliff, W., Bedoya, A.D., Balu, S., O'Brien, C., and Sendak, M.P. (2020). Integrating a Machine Learning System Into Clinical Workflows: Qualitative Study. J. Med. Internet Res. *22*, e22421.

21. Henry, K.E., Kornfield, R., Sridharan, A., Linton, R.C., Groh, C., Wang, T., Wu, A., Mutlu, B., and Saria, S. (2022). Human–machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. NPJ Digit. Med. *5*, 97. https://doi.org/10.1038/s41746-022-00597-7.

22. London, A.J. (2022). Artificial intelligence in medicine: Overcoming or recapitulating structural challenges to improving patient care? Cell Rep. Med. *3*, 100622.

23. Hey, S.P., and Kimmelman, J. (2013). Ethics, error, and initial trials of efficacy. Sci. Transl. Med. *5*, 184fs16.

24. Kimmelman, J. (2012). A theoretical framework for early human studies: uncertainty, intervention ensembles, and boundaries. Trials *13*, 173.

25. Kimmelman, J., and London, A.J. (2015). The structure of clinical translation: efficiency, information, and ethics. Hastings Cent. Rep. *45*, 27–39.

26. Zhou, Q., Chen, Z.-H., Cao, Y.-H., and Peng, S. (2021). Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. NPJ Digit. Med. *4*, 154.

27. Campbell, N.C., Murray, E., Darbyshire, J., Emery, J., Farmer, A., Griffiths, F., Guthrie, B., Lester, H., Wilson, P., and Kinmonth, A.L. (2007). Designing and evaluating complex interventions to improve health care. BMJ *334*, 455–459.

28. London, A.J., and Danks, D. (2018). Regulating Autonomous Vehicles. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. https://doi.org/10.1145/3278721.3278763.

29. Fraser, A.G., Biasin, E., Bijnens, B., Bruining, N., Caiani, E.G., Cobbaert, K., Davies, R.H., Gilbert, S.H., Hovestadt, L., Kamenjasevic, E., et al. (2023). Artificial intelligence in medical device software and high-risk medical devices - a review of definitions, expert recommendations and regulatory initiatives. Expet Rev. Med. Dev. *20*, 467–491.

30. Glenn Cohen, I., Minssen, T., Nicholson Price, W., II, Robertson, C., and Shachar, C. (2022). The Future of Medical Device Regulation: Innovation and Protection (Cambridge University Press).

31. Theisz, V. (2015). Medical Device Regulatory Practices: An International Perspective (CRC Press).

32. Papademetris, X., Quraishi, A.N., and Licholai, G.P. (2022). Introduction to Medical Software: Foundations for Digital Health, Devices, and Diagnostics (Cambridge University Press).

33. Liu, X., Cruz Rivera, S., Moher, D., Calvert, M.J., and Denniston, A.K.; SPIRIT-AI and CONSORT-AI Working Group (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Lancet. Digit. Health *2*, e537–e548.

34. Cruz Rivera, S., Liu, X., Chan, A.W., Denniston, A.K., and Calvert, M.J.; SPIRIT-AI and CONSORT-AI Working Group (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Lancet. Digit. Health *2*, e549–e560.

35. Vasey, B., Nagendran, M., Campbell, B., Clifton, D.A., Collins, G.S., Denaxas, S., Denniston, A.K., Faes, L., Geerts, B., Ibrahim, M., et al. (2022). Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nat. Med. *28*, 924–933.

36. Abràmoff, M.D., Lavin, P.T., Birch, M., Shah, N., and Folk, J.C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit. Med. *1*, 39.

37. Passi, S., and Barocas, S. (2019). Problem Formulation and Fairness. In Proceedings of the Conference on Fairness, Accountability, and Transparency (ACM). https://doi.org/10.1145/3287560.3287567.

38. Dhaliwal, G. (2020). Clinical Diagnosis—Is There Any Other Type? JAMA Intern. Med. *180*, 1304–1305. https://doi.org/10.1001/jamainternmed.2020.3048.

39. Keane, P.A., and Topol, E.J. (2018). With an eye to AI and autonomous diagnosis. NPJ Digit. Med. *1*, 40.

40. Shieh, Y., Eklund, M., Sawaya, G.F., Black, W.C., Kramer, B.S., and Esserman, L.J. (2016). Population-based screening for cancer: hope and hype. Nat. Rev. Clin. Oncol. *13*, 550–565.

41. Houssami, N., Given-Wilson, R., and Ciatto, S. (2009). Early detection of breast cancer: overview of the evidence on computer-aided detection in mammography screening. J. Med. Imaging Radiat. Oncol. *53*, 171–176.

42. Taylor, P., and Potts, H.W.W. (2008). Computer aids and human second reading as interventions in screening mammography: two systematic reviews to compare effects on cancer detection and recall rate. Eur. J. Cancer *44*, 798–807.

43. Topol, E.J. (2020). Welcoming new guidelines for AI clinical research. Nat. Med. *26*, 1318–1320.

44. Harvey, H., and Oakden-Rayner, L. (2020). Guidance for Interventional Trials Involving Artificial Intelligence. Radiol. Artif. Intell. *2*, e200228.

45. Aristidou, A., Jena, R., and Topol, E.J. (2022). Bridging the chasm between AI and clinical implementation. Lancet *399*, 620.

46. Thodberg, H.H., Thodberg, B., Ahlkvist, J., and Offiah, A.C. (2022). Autonomous artificial intelligence in pediatric radiology: the use and perception of BoneXpert for bone age assessment. Pediatr. Radiol. *52*, 1338–1346.

47. De, S.J., Guisez, T., Wijnand, J., Cools, M., Herregods, N., den Brinker, M., Gielen, J., Ernst, C., and Gies, I. (2021). The BoneXpert adult height prediction method outperforms the Bayley and Pinneau method in tall male adolescents. In 59th ESPE Annual Meeting (ESPE 2021 Online), *94* (European Society for Paediatric Endocrinology).

48. Adams, R., Henry, K.E., Sridharan, A., Soleimani, H., Zhan, A., Rawat, N., Johnson, L., Hager, D.N., Cosgrove, S.E., Markowski, A., et al. (2022). Prospective, multi-site study of patient outcomes after implementation of the TREWS machine learning-based early warning system for sepsis. Nat. Med. *28*, 1455–1460.

49. Martin, D.D., Calder, A.D., Ranke, M.B., Binder, G., and Thodberg, H.H. (2022). Accuracy and self-validation of automated bone age determination. Sci. Rep. *12*, 6388.

50. Ferryman, K. (2020). Addressing health disparities in the Food and Drug Administration's artificial intelligence and machine learning regulatory framework. J. Am. Med. Inf. Assoc. *27*, 2016–2019.

51. Vyas, D.A., Eisenstein, L.G., and Jones, D.S. (2021). Hidden in Plain Sight—Reconsidering the Use of Race Correction in Clinical Algorithms. Obstetrical & Gynecological Survey *76*, 5–7. https://doi.org/10.1097/01.ogx.0000725672.30764.f7.

52. McCradden, M.D., Odusi, O., Joshi, S., Akrout, I., Ndlovu, K., Glocker, B., Maicas, G., Liu, X., Mazwi, M., Garnett, T., Oakden-Rayner, L., et al. (2023). What's fair is… fair? Presenting JustEFAB, an ethical framework for operationalizing medical ethics and social justice in the integration of clinical machine learning: JustEFAB. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, pp. 1505–1519.

53. Tsiandoulas, K., McSheffrey, G., Fleming, L., Rawal, V., Fadel, M.P., Katzman, D.K., and McCradden, M.D. (2023). Ethical tensions in the treatment of youth with severe anorexia nervosa. Lancet. Child Adolesc. Health *7*, 69–76.

54. Pierson, E., Cutler, D.M., Leskovec, J., Mullainathan, S., and Obermeyer, Z. (2021). An algorithmic approach to reducing unexplained pain disparities in underserved populations. Nat. Med. *27*, 136–140.

55. DiMarco, M., Zhao, H., Boulicault, M., and Richardson, S.S. (2022). Why 'sex as a biological variable' conflicts with precision medicine initiatives. Cell Rep. Med. *3*, 100550.

56. Yearby, R. (2020). Structural Racism and Health Disparities: Reconfiguring the Social Determinants of Health Framework to Include the Root Cause. J. Law Med. Ethics *48*, 518–526.

57. Ray, K.S. (2021). It's Time for a Black Bioethics. Am. J. Bioeth. *21*, 38–40.

58. Mukwende, M., Tamony, P., and Turner, M. (2020). Mind the Gap: A Handbook of Clinical Signs in Black and Brown Skin.

59. McCradden, M.D., Joshi, S., Anderson, J.A., Mazwi, M., Goldenberg, A., and Zlotnik Shaul, R. (2020). Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. J. Am. Med. Inf. Assoc. *27*, 2024–2027.

60. Oza, C., Khadilkar, A.V., Mondkar, S., Gondhalekar, K., Ladkat, A., Shah, N., Lohiya, N., Prasad, H.K., Patil, P., Karguppikar, M., et al. (2022). A comparison of bone age assessments using automated and manual methods in children of Indian ethnicity. Pediatr. Radiol. *52*, 2188–2196.

61. Oakden-Rayner, L., Gale, W., Bonham, T.A., Lungren, M.P., Carneiro, G., Bradley, A.P., and Palmer, L.J. (2022). Validation and algorithmic audit of a deep learning system for the detection of proximal femoral fractures in patients in the emergency department: a diagnostic accuracy study. Lancet. Digit. Health *4*, e351–e358. https://doi.org/10.1016/s2589-7500(22)00004-8.

62. Oakden-Rayner, L., Dunnmon, J., Carneiro, G., and Ré, C. (2020). Hidden Stratification Causes Clinically Meaningful Failures in Machine Learning for Medical Imaging. Proc. ACM Conf. Health Inference Learn. *2020*, 151–159.

63. Futoma, J., Simons, M., Doshi-Velez, F., and Kamaleswaran, R. (2021). Generalization in Clinical Prediction Models: The Blessing and Curse of Measurement Indicator Variables. Crit. Care Explor. *3*, e0453.

64. Futoma, J., Simons, M., Panch, T., Doshi-Velez, F., and Celi, L.A. (2020). The myth of generalisability in clinical research and machine learning in health care. Lancet. Digit. Health *2*, e489–e492.

65. Park, Y., Jackson, G.P., Foreman, M.A., Gruen, D., Hu, J., and Das, A.K. (2020). Evaluating artificial intelligence in medicine: phases of clinical research. JAMIA Open *3*, 326–331.

66. Sendak, M., Vidal, D., Trujillo, S., Singh, K., Liu, X., and Balu, S. (2023). Surfacing Best Practices for AI Software Development and Integration in Healthcare (Frontiers Media SA).

67. Embi, P.J. (2021). Algorithmovigilance—Advancing Methods to Analyze and Monitor Artificial Intelligence–Driven Health Care for Effectiveness and Equity. JAMA Netw. Open *4*, e214622. https://doi.org/10.1001/jamanetworkopen.2021.

68. Finlayson, S.G., Subbaswamy, A., Singh, K., Bowers, J., Kupke, A., Zittrain, J., Kohane, I.S., and Saria, S. (2021). The Clinician and Dataset Shift in Artificial Intelligence. N. Engl. J. Med. *385*, 283–286.

69. Ensign, D., Friedler, S.A., Neville, S., Scheidegger, C., and Venkatasubramanian, S. (2018). Runaway Feedback Loops in Predictive Policing. In Conference on Fairness, Accountability and Transparency (PMLR), pp. 160–171.

70. Perdomo, J., Zrnic, T., Mendler-Dünner, C., and Hardt, M. (2020). Performative Prediction. In International Conference on Machine Learning (PMLR), pp. 7599–7609.

71. London, A.J. (2022). Overcoming or Recapitulating Fundamental Problems. Cell Reports Medicine *3*, 100622.

About the authors

Melissa D. McCradden is the John and Melinda Thompson Director of AI in Medicine, a bioethicist at The Hospital for Sick Children, and an associate scientist with the Genetics & Genome Biology Research Program at the Peter Gilgan Centre for Research and Learning (SickKids). She received her PhD in neuroscience from McMaster University and her master's in bioethics from the University of Toronto. Her research explores ethical issues for responsible evaluation of health technologies and how evidence contributes to clinical decision-making in pediatric healthcare. Melissa has published in *Nature Medicine*, *Lancet Digital Health*, *JAMA Pediatrics*, *JAMIA*, *PMLR*, and *CHIL*. She is a consensus group member for the CONSORT-AI, SPIRIT-AI, DECIDE-AI, QUADAS-AI, and STARD-AI reporting guidelines.

Shalmali Joshi is an assistant professor in the Department of Biomedical Informatics at Columbia University. She was previously a postdoctoral fellow at Harvard University and at the Vector Institute. She received her PhD from UT Austin. Her research is on the algorithmic safety of machine learning for healthcare. Shalmali has contributed to robustness, explainability, and developing novel algorithms for ML safety for healthcare. Shalmali has published in *NeurIPS*, *ICML*, *FAccT*, *CHIL*, *MLHC*, *PMLR*, *JAMIA*, *LDH*, and *Nature Medicine*, co-founded the Fair ML for Health NeurIPS workshop, and has served as the general chair of ML4H 2022 and the program chair of MLHC 2023.

James A. Anderson, MHA, MA, PhD, is bioethicist in the Department of Bioethics, The Hospital for Sick Children; a project investigator at the SickKids Research Institute; an assistant professor at the Institute for Health Policy, Management, and Evaluation at the University of Toronto; and a member of the University's Joint Centre for Bioethics. His research explores the intersection of ethics and epistemology in clinical research and practice. Recent work has focused on ML applications in healthcare, whole-genome sequencing, and clinical decision-making in pediatrics.

Alex John London is the K&L Gates Professor of Ethics and Computational Technologies at Carnegie Mellon University, where he directs the Center for Ethics and Policy. An elected fellow of the Hastings Center, his work deals with ethical issues in biomedical research, artificial intelligence, and normative ethics. In 2022, his book "For the Common Good: Philosophical Foundations of Research Ethics" was published by Oxford University Press, and he is the author of over 100 papers or book chapters that have appeared in *Mind*, *Science*, *The Lancet*, *JAMA*, *Statistics in Medicine*, *The Philosopher's Imprint*, the *Hastings Center Report*, and many others.