frontiers | Research Topics

Surfacing best practices for AI software development and integration in healthcare

Edited by
Mark Sendak, Suresh Balu, Xiao Liu, Karandeep Singh, Sylvia Trujillo and David Vidal

Published in
Frontiers in Digital Health

ISSN 1664-8714
ISBN 978-2-8325-2580-7
DOI 10.3389/978-2-8325-2580-7

About Frontiers

Frontiers is more than just an open access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers journal series

The Frontiers journal series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the *Frontiers journal series* operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews. Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the *Frontiers journals series*: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area.

Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers editorial office: frontiersin.org/about/contact

Surfacing best practices for AI software development and integration in healthcare

Topic editors
Mark Sendak — Duke University, United States
Suresh Balu — Duke University, United States
Xiao Liu — University Hospitals Birmingham NHS Foundation Trust, United Kingdom
Karandeep Singh — University of Michigan, United States
Sylvia Trujillo —Sylvia J Trujillo, United States
David Vidal — Mayo Clinic, United States

Citation
Sendak, M., Balu, S., Liu, X., Singh, K., Trujillo, S., Vidal, D., eds. (2023). *Surfacing best practices for AI software development and integration in healthcare*.
Lausanne: Frontiers Media SA. doi: 10.3389/978-2-8325-2580-7

Dr. Karandeep Singh has received grant funding from Blue Cross Blue Shield of Michigan and Teva Pharmaceuticals. All other Topic Editors declare no conflicts of interest.

Table of contents

TYPE Editorial
PUBLISHED 21 February 2023
DOI 10.3389/fdgth.2023.1150875

OPEN ACCESS

EDITED AND REVIEWED BY
Daniel B. Hier,
Missouri University of Science and Technology,
United States

*CORRESPONDENCE
Mark Sendak
✉ mark.sendak@duke.edu

SPECIALTY SECTION
This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health

RECEIVED 25 January 2023
ACCEPTED 06 February 2023
PUBLISHED 21 February 2023

CITATION
Sendak M, Vidal D, Trujillo S, Singh K, Liu X and Balu S (2023) Editorial: Surfacing best practices for AI software development and integration in healthcare.
Front. Digit. Health 5:1150875.
doi: 10.3389/fdgth.2023.1150875

Editorial: Surfacing best practices for AI software development and integration in healthcare

Mark Sendak[1]*, David Vidal[2], Sylvia Trujillo[3], Karandeep Singh[4], Xiaoxuan Liu[5] and Suresh Balu[1]

[1]Duke Institute for Health Innovation, Durham, NC, United States, [2]Mayo Clinic, Rochester, MN, United States, [3]OCHIN, Portland, OR, United States, [4]Division of Nephrology, Department of Internal Medicine, University of Michigan, Ann Arbor, MI, United States, [5]Academic Unit of Ophthalmology, Institute of Inflammation and Ageing, University of Birmingham, Birmingham, United Kingdom

KEYWORDS
health innovation, health policy, technology regulation, technology diffusion, machine learning, artificial intelligence

Editorial on the Research Topic
Surfacing best practices for AI software development and integration in healthcare

Introduction

The evidence supporting the mainstream use of artificial intelligence (AI) software in healthcare is rapidly mounting. Three systematic reviews of AI software randomized controlled trials (RCTs) were published in 2021 and 2022, including 95 studies across 29 countries (1–3). In the United States (US), the Centers for Medicare and Medicaid Services (CMS) is approving AI software systems for reimbursement through multiple payment mechanisms (4). In the United Kingdom (UK), the National Screening Committee is exploring the use of AI software in national cancer screening and has awarded £90 million to prospective multi-center trials (5, 6). Two large, multi-hospital studies showed the mortality benefit of early detection and treatment of inpatient deterioration and pneumonia (7, 8). However, despite advances in technology and policy, isolated success stories are not leading to efficient diffusion of validated AI software across settings.

A key barrier preventing efficient translation of AI software to new clinical settings is the lack of visibility into poorly characterized, yet critically important labor that Mary Gray and Siddharth Suri call “Ghost Work” (9). Ghost work is broadly described as the invisible labor that powers technology platforms. In healthcare, ghost work is carried out by front-line clinical and administrative staff working beyond the contours of technical AI systems to effectively integrate the technologies into local social environments. But while the brittleness of AI software systems over time and across sites is broadly recognized (10, 11), health systems develop strategies largely in silos. To fill this gap, we invited teams from health systems around the globe to contribute to the research topic “Surfacing Best Practices for AI Software Development and Integration in Healthcare (12).” The research topic was sponsored by Janssen Pharmaceuticals of Johnson & Johnson. In this editorial, we present a synthesis of the nine featured manuscripts and highlight strategies used across settings as well as future opportunities for development and partnership.

Methods

We conducted two primary analyses of the nine research topic manuscripts to identify key themes. We then complement the two primary analyses with details about the host institution, country, model use case, manuscript objectives, and key takeaways.

In the first primary analysis, we mapped the topics described in each manuscript to various stages of the AI software lifecycle. The four stages are defined as follows. First, *problem definition and solution procurement* describes the activities related to how organizations identify and prioritize problems and then allocate resources and personnel to pursue opportunities. Second, *AI solution development and adaptation* describes the activities related to how organizations either build technologies internally or adapt externally built tools. Third, *technical and clinical integration* describes the activities related to how organizations integrate AI solutions into legacy information technology systems and clinical workflows, roles, and responsibilities. Fourth, *lifecycle management* describes the activities related to maintenance, updating, and decommissioning of AI solutions used in clinical care. Each research topic manuscript could be mapped to multiple lifecycle stages.

In the second primary analysis, we reviewed biosketches, organization websites, and professional social media pages to map each research topic manuscript author to formal academic training across disciplines. Due to the large number of manuscript authors and broad range of formal training, we grouped disciplines into seven categories: engineering, computer science, and physics; statistics, biostatistics, and bioinformatics; business and management; public health and economics; biological or behavioral science; clinical doctorate; ethics or bioethics. Each author could be mapped to multiple academic disciplines.

Results

The research topic "Surfacing Best Practices for AI Software Development and Integration in Healthcare" features 9 manuscripts with 73 authors from 7 institutions across 4 countries. Two institutions published two manuscripts each, including The Hospital for Sick Children in Toronto, Canada and University of Wisconsin in Madison, Wisconsin, USA. The AI software use cases featured in the research topic include three pediatric applications (hydronephrosis due to obstruction, arrhythmia detection, and sleep-wake patterns in neonates), one mental health application (suicide prevention), three general adult applications (30-day readmission, inpatient deterioration, and new-onset atrial fibrillation), and two geriatrics applications (advance care planning, falls risk in the emergency department). One research topic manuscript describes an organizational governance framework that has overseen ten AI software integrations, two decommissions, and one decision to not integrate (Liao et al.). Additional information about the use cases and key takeaways are presented in **Table 1**.

AI software lifecycle stages

The research topic features manuscripts that contribute insights related to all four AI software lifecycle stages (problem definition and solution procurement, development and adaptation, technical and clinical integration, and lifecycle management). Two manuscripts describe programs that span all lifecycle stages, including the implementation of an AI quality management system at University Medical Center in Ultrecht, Netherlands and an AI organizational governance process at University of Wisconsin in Madison, USA. Two manuscripts present different frameworks for AI solution development, technical and clinical integration, and lifecycle management. A team from The Hospital for Sick Children in Toronto, Canada presents an approach that adopts language from systems engineering, while a team from University College London in the UK presents an approach that adopts language from therapeutics development (Assadi et al.). Two manuscripts present case studies focused on technical and clinical integration, including an adult deterioration model integrated at St. Michael's Hospital in Toronto, Canada, and a falls risk model integrated at University of Wisconsin in Madison, USA (Pou-Prom et al.). Lastly, three manuscripts present best practices related to specific lifecycle stages. A team from The Hospital for Sick Children in Toronto, Canada describes the use of AI software silent trials during technical integration (Kwong et al.), a team from Stanford Health Care in Palo Alto, USA describes reliability and fairness audits during lifecycle management (Lu et al.), and a team from Vanderbilt Health describes AI solution monitoring and updating during lifecycle management (Davis et al.).

Team composition

In some ways, the research topic authorship teams are similar. All manuscripts feature interdisciplinary teams at academic health centers and graduate students and clinical trainees made significant contributions as co-authors. All manuscripts include clinical and technical expert co-authors. And lastly, all manuscripts build on prior work from authorship teams who have previously published AI solution validation studies.

In other ways, the research topic authorship teams are heterogeneous. The smallest teams were a group of three clinicians and informaticians at Vanderbilt Health who describe AI software monitoring and updating challenges and a group of four engineers, public health experts, and clinicians who describe the AI software organizational governance model at University of Wisconsin. The largest team was a group of twenty-seven

TABLE 1 Summary of key information from each research topic manuscript.

Name	Institution	Country	Model use case	Manuscript objective	AI translation phase	Key takeaways	Author domain expertise
An integration engineering framework for machine learning in healthcare	The Hospital for Sick Children	Canada	• "We have applied this framework to an arrhythmia detection model and have implemented this as a best practice at the Hospital for Sick Children. This practical application is demonstrated in the supplementary material."	"Improper integration of new systems may lead to additional costs, patient harm, damage to other systems, and decrease in efficiency. To address this translation gap, we present a systems engineering framework to guide the development of models with explicit consideration of elements that are crucial for successful model integration in healthcare."	Development; Technical Integration; Lifecycle management	• Applies systems engineering to the process of integrating machine learning in healthcare, describing the domains of integration (e.g., the technical system, human, and environment) and the interactions between the domains. • Conducted a narrative review to understand the challenges and gaps in integrating ML in health care, challenges associated with the current software development life cycle, and principles of integration engineering. • Present a generalizable framework for ML integration in health care with four phases: 1) inception, 2) preparation, 3) development, 4) integration.	Engineering, CS, or Physics: 8 Statistics, Biostatistics, or Bioinformatics: 0 Business or Management: 0 Public Health or Economics: 1 Biological or Behavioral sciences: 2 Clinical Doctorate: 6 Ethics or Bioethics: 1 **Total number of authors: 13** **Total number of domains: 5**
Clinical deployment environments: Five pillars of translational machine learning for health	University College London Hospital	UK	• "Our algorithm evaluates the risk of new onset atrial fibrillation (NOAF) in real-time based on existing electrolyte levels, medications, disease type, and the patient's co-morbidities… Our model is now used to drive a CDSS that operates in two layers. Firstly, where electrolytes are outside existing (evidence based guidelines) the CDSS makes a strong deterministic recommendation. Secondly, where electrolytes are within the window of the broader guideline but could be optimised, the CDSS makes a nudged randomised recommendation based on the model's prediction."	"In this paper, we describe the functional requirements for a Clinical Deployment Environment (CDE) for translational ML4H. These requirements map closely to the classical components of translational medicine, but differ in that algorithms will require ongoing stewardship even after a successful deployment. The CDE is an infrastructure that manages algorithms with the same regard that is given to medicines (pharmacy) and machines (medical physics)."	Development; Technical Integration; Lifecycle management	• Presents five pillars to a clinical deployment environment for translating machine learning for health: 1) real world development; 2) ML-ops for health; 3) responsible AI in practice; 4) implementation science; 5) continuous evaluation. • Describes the similarities between translation of machine learning into healthcare and drug development. • Describes in great detail (in the supplement) the Experimental Medicine Application Platform (EMAP), where ML researchers iteratively build, validate, and test models, and the FlowEHR ML-Ops platform, which supports the deployment and maintenance of local models.	Engineering, CS, or Physics: 4 Statistics, Biostatistics, or Bioinformatics: 0 Business or Management: 0 Public Health or Economics: 1 Biological or Behavioral sciences: 0 Clinical Doctorate: 2 Ethics or Bioethics: 0 **Total number of authors: 6** **Total number of domains: 3**

(*continued*)

TABLE 1 Continued

Name	Institution	Country	Model use case	Manuscript objective	AI translation phase	Key takeaways	Author domain expertise
The silent trial—the bridge between bench-to-bedside clinical AI applications	The Hospital for Sick Children	Canada	• "Classification model to predict obstruction in hydronephrotic kidneys of infants using ultrasound images" • "Develop an AI model that could reliably distinguish between self-resolving hydronephrosis vs. those that would ultimately require operative management based on initial kidney ultrasound"	"The purpose of this article is to highlight the lessons learned from our experience in validating a previously developed model within the context of the silent trial… Using our model as a case study, we illustrate how issues related to dataset drift, bias, feasibility, and stakeholder attitudes were identified and addressed. This article is intended for clinicians and ML engineers wishing to gain a deeper understanding of the rationale behind the silent trial and provide insights as to how this phase serves as a bridge between initial model development and clinical trials assessment."	Technical Integration; Clinical Integration	• Describes a 2-step silent trial where first step is to assess generalization prospectively and second step retrained model and re-evaluated performance prospectively. • Describes approach to update model after silent trial to mitigate effect of dataset shift • Assessed patient and family perceptions about AI with post-visit follow-up questionnaire	Engineering, CS, or Physics: 3 Statistics, Biostatistics, or Bioinformatics: 2 Business or Management: 0 Public Health or Economics: 1 Biological or Behavioral sciences: 1 Clinical Doctorate: 3 Ethics or Bioethics: 1 **Total number of authors: 8** **Total number of domains: 6**
Operationalizing a real-time scoring model to predict fall risk among older adults in the emergency department	University of Wisconsin Health	USA	• "Our research team has developed and validated an innovative automated screening algorithm that uses machine learning coupled with electronic health record (EHR) data to predict fall risk in the 180 days following an ED visit using retrospective data (14)." • "This algorithm had the promise of identifying older adults at high risk of falling in the 6 months following the ED visit. Furthermore, engaging with experts in human factors engineering and clinicians, the study team designed a workflow and alerts designed to create a system in which the algorithm facilitates screening of older adult patients in the ED and facilitating referral for fall prevention services (15)."	"This case-study describes challenges and barriers we overcame in the use of such a model after it had been created and validated in silico. Based on this experience, we provide general principles for translating an EHR-based predictive model from research and reporting environments into real-time operation."	Technical integration; Clinical Integration	• Detailed description of how the team made modifications to an algorithm through three stages over 15 months, including: stage 1) training and testing the algorithm on a research dataset; stage 2) validating the algorithm on production-system data; stage 3) live implementation of the algorithm. • Discussion of several unexpected technical challenges, including IT constraints to operationalize the model, model interpretability, model threshold selection, and model placement in the workflow.	Engineering, CS, or Physics: 2 Statistics, Biostatistics, or Bioinformatics: 0 Business or Management: 0 Public Health or Economics: 2 Biological or Behavioral sciences: 1 Clinical Doctorate: 1 Ethics or Bioethics: 0 **Total number of authors: 5** **Total number of domains: 4**

(*continued*)

TABLE 1 Continued

Name	Institution	Country	Model use case	Manuscript objective	AI translation phase	Key takeaways	Author domain expertise
From compute to care: Lessons learned from deploying an early warning system into clinical practice	St Michael's Hospital	Canada	• "In Fall 2020, we deployed CHARTwatch to the General Internal Medicine (GIM) ward at St. Michael's Hospital, an inner-city teaching hospital in Canada." • "We developed a model [CHARTwatch] to detect inpatient deterioration, defined as in-hospital death or transfer to the intensive care unit (ICU)."	"Here, we describe in detail, the system's infrastructure and assess the success of our deployment through quantitative metrics (such as model performance, end-user engagement, and adherence to workflows) and by comparing our deployment to the GMLP principles. The purpose of this manuscript is to provide concrete insights into the deployment of ML in a healthcare setting and highlight opportunities to strengthen GMLP guidance."	Technical Integration; Clinical Integration	• Detailed description of the approach taken to minimize alert fatigue, including a 48 h snooze, 24 h snooze after ICU discharge, and silencing alerts after the fifth instance. • Description of changes made in response to a silent trial, including adapting model for high sensitivity troponin. • Presents best practices for downtime protocols (e.g., emails to IT team if script fails), end-user engagement, and training. • Presents concrete descriptions of how the team operationalized the 10 GMLP recommendations	Engineering, CS, or Physics: 1 Statistics, Biostatistics, or Bioinformatics: 1 Business or Management: 1 Public Health or Economics: 2 Biological or Behavioral sciences: 0 Clinical Doctorate: 2 Ethics or Bioethics: 0 **Total number of authors: 5** **Total number of domains: 5**
Considerations in the reliability and fairness audits of predictive models for advance care planning	Stanford Health Care	USA	• "In this work, we illustrate a reliability/fairness audit of 12-month mortality models considered for use in supporting team-based advance care planning (ACP) in three practice settings."	"We (1) design and report a reliability/fairness audit of the models following existing reporting guidelines, (2) survey decision makers about how the results impacted their decision of whether to use the model, and (3) quantify the time, workflow and data requirements for performing this audit. We discuss key drivers and barriers to making these audits standard practice. We believe this may aid other decision makers and informaticists in operationalizing regular reliability and fairness audits."	Lifecycle management	• Assessed two models, including an Epic end-of-life index and a homegrown 12-month mortality model, by performing a reliability audit (model performance and calibration) and fairness audit (summary statistics, subgroup performance, subgroup calibration). • Present audit results for both algorithms and determined differences in model performance and calibration across demographic subgroups. • Presents results of a survey sent to decision makers to understand the role of fairness and robustness audits in algorithmic governance. • Discuss the resource requirements and amount of effort (115 h) required to conduct the audit.	Engineering, CS, or Physics: 5 Statistics, Biostatistics, or Bioinformatics: 5 Business or Management: 1 Public Health or Economics: 3 Biological or Behavioral sciences: 2 Clinical Doctorate: 14 Ethics or Bioethics: 0 **Total number of authors: 27** **Total number of domains: 6**

(*continued*)

TABLE 1 Continued

Name	Institution	Country	Model use case	Manuscript objective	AI translation phase	Key takeaways	Author domain expertise
Open questions and research gaps for monitoring and updating AI-enabled tools in clinical settings	Vanderbilt University Medical Center	USA	• "To explore performance drift in an operational setting, we evaluated the performance of two models currently implemented in the production EHR system at Vanderbilt University Medical Center (VUMC): a non-proprietary, externally developed model predicting readmission (LACE+) (23) and a locally developed model predicting suicidal behaviors (Vanderbilt Suicide Attempt and Ideation Likelihood model, VSAIL) (24)."	"In this paper, we highlight the need for maintaining clinical prediction models and discuss open questions regarding this critical aspect of the AI modeling lifecycle. First, we illustrate performance drift across models implemented in the production electronic health record (EHR) system at an academic medical center. Second, we discuss several open research questions and describe the nuances required for best practice guidance."	Lifecycle management	• Presented two different types of model performance drift patterns using observed to expected outcome ratio (O:E) that occurred for a non-proprietary 30-day readmission model (LACE+) and for a homegrown suicide prevention model (VSAIL). • Detailed discussion of important questions related to three areas of lifecycle management: model maintenance policies (e.g., how should model ownership impact local control over maintenance?), performance monitoring perspectives (e.g., at what level should model performance be maintained? what aspects of performance should be monitored?), and model updating strategies (e.g., what updating approaches should be considered?).	Engineering, CS, or Physics: 0 Statistics, Biostatistics, or Bioinformatics: 3 Business or Management: 0 Public Health or Economics: 1 Biological or Behavioral sciences: 0 Clinical Doctorate: 2 Ethics or Bioethics: 0 **Total number of authors: 3** **Total number of domains: 3**
Governance of Clinical AI applications to facilitate safe and equitable deployment in a large health system: Key elements and early successes	University of Wisconsin Health	USA	• "At the time of this publication, the governance framework has overseen ten successful deployments, two successful retirements, and one successful non-deployment across nine applications." • "Applications include diverse use of AI prediction for outputs including severe sepsis, clinical deterioration, physician panel weighting, COVID detection on radiographs, emergency department screening for falls prevention, screening for opioid abuse, and ED crowding"	"As we expand our technical ability to provide solutions, more skepticism and questions surface, and at times resistance, around the suitability of using AI in routine clinical care from all levels of the organization, ranging from front-line clinical staff to executive leadership. In response to these questions and the challenges for implementation, the health system recognized the need for a governance structure to endorse and oversee adoption, implementation, and ongoing value evaluation of AI-driven applications. This case study describes the development and nature of governance of clinical AI applications at our institution."	Full lifecycle	• Describe clinical, operational, and leadership challenges encountered when establishing an institutional governance process. • Describes creation of institutional-level "Clinical AI and Predictive Analytics Committee" that oversees work conducted by use-case specific algorithm workgroups. • Presents five guiding principles that have emerged for the governance committee. • Describes approach for ongoing AI model monitoring, including a successful model retirement, and mechanisms to incorporate equity and ethics concerns related to AI	Engineering, CS, or Physics: 2 Statistics, Biostatistics, or Bioinformatics: 0 Business or Management: 0 Public Health or Economics: 3 Biological or Behavioral sciences: 0 Clinical Doctorate: 2 Ethics or Bioethics: 0 **Total number of authors: 4** **Total number of domains: 3**

(*continued*)

TABLE 1 Continued

Name	Institution	Country	Model use case	Manuscript objective	AI translation phase	Key takeaways	Author domain expertise
A Perspective on a Quality Management System for AI/ML-Based Clinical Decision Support in Hospital Care	University Medical Center Utrecht	Netherlands	• "Sleep Well Baby…is an in-house developed ML model intended for monitoring real-time sleep-wake patterns in preterm neonates between 28- and 34-weeks gestational age." • "The added value of real-time sleep-wake state monitoring comes from adapting elective clinical management of these preterm infants toward less disturbance during sleep periods."	"In this perspective we illustrate our learnings regarding quality management of AI/ML-CDS tools through an example from our development pipeline, Sleep Well Baby (SWB). After introducing the SWB project and describing the development phase we address life-cycle management questions that arose while operationalizing SWB. When addressing these questions, we illustrate how the organizational structure of medical laboratories and ISO15189 can inspire healthcare institutes in building an effective and sustainable Quality Management System (QMS) for AI/ML usage in clinical care. Finally, in the discussion we provide an outlook how quality management of AI/ML-CDS extends to third-party AI/ML tools and settings outside healthcare institutes other than academic teaching hospitals."	Full lifecycle	• Describes an innovation funnel process geared towards AI/ML product development process that became the blueprint for a national AI innovation tool. The funnel is divided into seven distinct phases with transition gates and references relevant EU-laws and regulations, guidelines, and standards. • Detailed description of how quality standards including IEC 62304 and ISO 14971 were applied during the development process of SWB. • Presents responses to highly relevant, practical life-cycle management questions: 1) Who is responsible for the AI/ML CDS device configuration; 2) Who gives clearance for the use of SWB in clinical practice?; 3) How to ensure safe change management and revision of SWB?; 4) What if model performance starts degrading?; 5) Who provides a helpdesk for users; 6) How are users trained?	Engineering, CS, or Physics: 1 Statistics, Biostatistics, or Bioinformatics: 1 Business or Management: 0 Public Health or Economics: 0 Biological or Behavioral sciences: 3 Clinical Doctorate: 3 Ethics or Bioethics: 0 **Total number of authors: 5** **Total number of domains: 4**

engineers, bioinformaticians, managers, public health experts, biological science experts, and clinicians who conducted fairness and robustness audits of multiple models at Stanford Health Care. All teams included experts with formal training in at least three of the disciplines listed in **Table 1** and two teams included experts with formal training in six disciplines. Among the 73 authors who contributed to the research topic, two perspectives were unique. There was a single AI ethics expert from The Hospital for Sick Children in Toronto, Canada and there was a senior data scientist at University Medical Center in Utrecht, Netherlands who is also a clinical microbiologist who has implemented and audited laboratory quality management systems.

Discussion

The research topic "Surfacing Best Practices for AI Software Development and Integration in Healthcare" features a remarkably diverse set of insights and learnings from teams around the globe integrating and using AI software into practice (12). Throughout the research topic, teams consistently describe responses to unexpected challenges encountered in the transition from conducting AI software research to translating a technology into practice. The success of AI in healthcare hinges on the ability to adapt and transition from research into routine clinical practice. Sharing challenges, failures and describing promising approaches that were implemented in real-world settings can inform teams around the globe looking to advance the use of AI software in healthcare.

Across the research topic, consensus emerged around three important AI software integration practices. First, many teams highlighted the importance of simulating AI software performance in local, operational settings prior to initial use in clinical care. One method discussed in multiple articles involved the operationalization of a "silent trial," during which bedside clinicians are initially blinded to the AI software as it is prospectively applied on operational data. While not novel, consensus is emerging around the importance of this activity (13–15). Silent trials can alert AI software developers to potential patient safety risks, bias, or integration concerns prior to clinical testing in a manner that minimizes risk to patients. Another article described the creation of a synthetic clinical deployment environment that anticipates real-world clinical decision making (Harris et al.).

Second, many teams highlighted the importance of AI software governance and management. Articles highlighted the importance of transdisciplinary teams and the need to assign responsibility and accountability to oversee AI software performance and appropriate use. One team used international standards to create a quality management system for AI software lifecycle management (Bartels et al.). Manuscripts in the research topic build upon existing frameworks and broaden the focus from AI software manufacturers to humans within health systems who oversee AI software used in clinical settings. The frameworks complement national efforts to equip the healthcare workforce to effectively adopt AI (16).

Lastly, many teams highlighted the importance of ongoing AI software monitoring and auditing. Some articles used existing standards for evaluating AI, including Health Canada/FDA/MHRA Joint Statement on 10 guiding principles for Good Machine Learning Practices (GMLP), however real-world experience led to additional recommendations, such as emphasizing user engagement, utilizing a silent trial, and creating downtime protocols. Another team described periodic reliability and fairness audits that went beyond quantitative comparison of AI software performance across demographic subgroups to also include stakeholder interviews to better understand the impact of the AI software.

While consensus emerged on the themes described above, the research topic did surface divergent perspectives on the importance of interpretability and explainability of AI software. For example, the teams at University of Wisconsin and University College London explicitly promote the use of explainable models. One team explained that "a desire to ensure we had an interpretable model further influenced our choice to pursue regression rather than tree-based models (Engstrom et al.)." The other team explained that "most AI models that operate as "black-box models" are unsuitable for mission-critical domains, such as healthcare, because they pose risk scenarios where problems that occur can remain masked and therefore undetectable and unfixable" (Harris et al.). This perspective offers a contrasting view from prior work examining the use of "black-box models" in clinical care (17), the limitations of current explainability methods (18), and the approach of regulators at the U.S. Food and Drug Administration (19). The research topic exposes the urgent need for research and policies that help organizations understand whether or not to prioritize AI software interpretability and explainability.

Future directions

The research topic reveals five important opportunities to advance AI software integration in health care, summarized in Box 1. First, governments and health systems must invest in building and sustaining transdisciplinary teams that manage AI software integrations. Best practices did not emerge from the heroic acts of individual scientists, but rather from transdisciplinary teams of experts working with health systems. These types of roles are often funded through health system operations and require significant investment.

Second, health systems must broaden stakeholder engagement throughout the AI software lifecycle. Unfortunately, only a single instance of direct patient engagement was described in the research topic, occurring at The Hospital for Sick Children in Toronto, Canada. Otherwise, there was limited patient and community engagement. And while the research topic authors were diverse, there was minimal representation of legal and regulatory experts and social scientists. These perspectives are crucial to ensure that AI software integration aligns with

BOX 1 Five recommendations that emerged from research topic manuscripts

1) Governments and health systems must invest in transdisciplinary teams that manage AI software integrations
2) Health systems must broaden stakeholder engagement to include patients, legal and regulatory experts, and social scientists
3) Practitioner and research community must standardize AI software integration definitions, processes, and procedures, as well as communication approaches
4) Governments and health systems must establish durable, multi-stakeholder collaboratives to continue surfacing and disseminating AI software integration best practices
5) Governments must fund programs designed to foster the adoption of well-validated AI software beyond highly resourced academic health systems

rapidly evolving regulations, and unintended consequences of AI software integration and use are anticipated, identified, and mitigated.

Third, there is an urgent need to develop and formalize standard AI software integration definitions, processes, and procedures as well as communication approaches (20). The research topic features teams that used language from different disciplines to describe AI software integration, including drug discovery, systems engineering, and international quality management standards. While it's important to build upon existing work across disciplines, the multiplicity of terms creates unnecessary ambiguity and confusion. Precisely defined steps and procedures need to be specified for rapid diffusion of more mature best practices, such as the "silent trial".

Fourth, durable, multi-stakeholder collaboratives are needed to continue surfacing and disseminating AI software integration best practices. Efforts that we are directly involved in to achieve this aim are the Health AI Partnership (21) to disseminate best practices across health systems and the development of AI software reporting standards, including DECIDE-AI (22), CONSORT-AI (23), STARD-AI (24), and SPIRIT-AI (25).

Fifth, the research topic highlights the importance of fostering the adoption of well-validated AI software beyond highly resourced academic health systems. Persistence of the status quo, where AI software is best integrated within settings with the most expertise, will undermine the potential benefit of AI software. Business models and public sector programs must be designed to enable academic health systems to support smaller under-resourced settings that do not have the internal capabilities to utilize AI software most effectively. One research topic manuscript described a promising approach: "For smaller entities, such as a single general practitioner, this effort [to establish an AI software quality management system] seems unfeasible. In this situation, complete dependence on the manufacturer is imaginable, making it difficult to establish truly safe performance. Again, inspiration can be found in the regional services of medical laboratories that very often provide access to competences and resources for safe application of diagnostics. Regional AI labs could provide services for the development, acquisition, and quality control of AI/ML for smaller healthcare institutes including general practitioners (Bartels et al.)." Programs that test different approaches of regional, multi-institutional support are urgently needed to ensure equitable diffusion of AI software.

Conclusion

The research topic "Surfacing Best Practices for AI Software Development and Integration in Healthcare" successfully surfaced best practices from 7 organizations across 4 countries. All teams were based at academic health systems and had previously published AI software validation studies. The research topic features insights across the AI software integration lifecycle and contributing authors represent diverse domains of expertise. There was consensus around the importance of local evaluations of AI software in a "silent trial", establishing organizational governance structures for AI software, and monitoring of technologies post-integration. However, the research topic also exposed limitations of current work and we present five recommendations to further advance AI software integration across settings. We hope our work informs AI software developers and policy makers and contributes to future efforts to broadly engage stakeholders in multi-institutional learning collaboratives.

Author contributions

MPS and DV wrote the first draft. All authors contributed to both the subsequent drafting and critical revision of the manuscript. All authors contributed to the article and approved the submitted version.

Conflict of interest

MPS and SB are co-inventors of intellectual property licensed by Duke University to Clinetic, Inc. and Cohere-Med, Inc. MPS and SB hold equity in Clinetic, Inc. MPS and SB receive funding from the Gordon and Betty Moore Foundation, Patrick J McGovern Foundation, and NIH. KS's institution receives grant funding from Teva Pharmaceuticals and Blue Cross Blue Shield of Michigan for unrelated work, and KS serves on an advisory board for Flatiron Health. XL receives funding from the Wellcome Trust, the National Institute of Health Research/NHSX/Health Foundation, the Alan Turing Institute, the MHRA, and NICE.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Zhou Q, Chen ZH, Cao YH, Peng S. Clinical impact and quality of randomized controlled trials involving interventions evaluating artificial intelligence prediction tools: a systematic review. *Npj Digit Med.* (2021) 4(1):154. doi: 10.1038/s41746-021-00524-2

2. Plana D, Shung DL, Grimshaw AA, Saraf A, Sung JJY, Kann BH. Randomized clinical trials of machine learning interventions in health care. *JAMA Netw Open.* (2022) 5(9):e2233946. doi: 10.1001/jamanetworkopen.2022.33946

3. Lam TYT, Cheung MFK, Munro YL, Lim KM, Shung D, Sung JJY. Randomized controlled trials of artificial intelligence in clinical practice: systematic review. *J Med Internet Res.* (2022) 24(8):e37188. doi: 10.2196/37188

4. Parikh RB, Helmchen LA. Paying for artificial intelligence in medicine. *Npj Digit Med.* (2022) 5(1):63. doi: 10.1038/s41746-022-00609-6

5. Taylor-Phillips S, Seedat F, Kijauskaite G, Marshall J, Halligan S, Hyde C, et al. UK National screening committee's approach to reviewing evidence on artificial intelligence in breast cancer screening. *Lancet Digital Heal.* (2022) 4(7):e558–65. doi: 10.1016/S2589-7500(22)00088-7

6. The Artificial Intelligence in Health and Care Award. Available at: https://transform.england.nhs.uk/ai-lab/ai-lab-programmes/ai-health-and-care-award/ (Accessed January 21, 2023).

7. Dean NC, Vines CG, Carr JR, Rubin JG, Webb BJ, Jacobs JR, et al. A pragmatic stepped-wedge, cluster-controlled trial of real-time pneumonia clinical decision support. *Am J Resp Crit Care.* (2022) 205(11):1330–6. doi: 10.1164/rccm.202109-2092OC

8. Escobar GJ, Liu VX, Schuler A, Lawson B, Greene JD, Kipnis P. Automated identification of adults at risk for in-hospital clinical deterioration. *N Engl J Med.* (2020 Nov 12) 383(20):1951–60. doi: 10.1056/NEJMsa2001090

9. Gray M, Suri S. *Ghost work.* New York, United States: Harper Business (2019).

10. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. *New Engl J Med.* (2021) 385 (3):283–6. doi: 10.1056/NEJMc2104626

11. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med.* (2021) 181(8):1065–70 . doi: 10.1001/jamainternmed.2021.2626

12. Surfacing Best Practices for AI Software Development and Integration in Healthcare. Available at: https://www.frontiersin.org/research-topics/28021/surfacing-best-practices-for-ai-software-development-and-integration-in-healthcare (Accessed January 21, 2023).

13. McCradden MD, Anderson JA, Stephenson EA, Drysdale E, Erdman L, Goldenberg A, et al. A research ethics framework for the clinical translation of healthcare machine learning. *Am J Bioeth.* (2022) 22(5):8–22. doi: 10.1080/15265161.2021.2013977

14. Wiens J, Saria S, Sendak MP, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med.* (2019) 25:1337–40. doi: 10.1038/s41591-019-0548-6

15. Sendak MP, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. Real-World integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR Med Inform.* (2020) 8(7):e15182. doi: 10.2196/15182

16. Horizon Scanning. Available at: https://digital-transformation.hee.nhs.uk/building-a-digital-workforce/dart-ed/horizon-scanning (Accessed January 21, 2023).

17. Sendak MP, Elish MC, Gao M, Futoma J, Ratliff W, Nichols M, et al. *The human body is a black box": supporting clinical decision-making with deep learning. FAT* '20: conference on fairness, accountability, and transparency; vol. 44* (2020). p. 99–109. Available from: https://dl.acm.org/doi/pdf/10.1145/3351095.3372827?download=true

18. Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health.* (2021) 3(11): e745–50. doi: 10.1016/S2589-7500(21)00208-9

19. Ross C. A "disaster", or a "clear path" forward?: New FDA guidance on AI in medicine sparks strong reactions. STAT News. (2022). Available from: https://www.statnews.com/2022/09/28/fda-artificial-intelligence-tools-regulation-oversight/ (Accessed January 21, 2023).

20. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. *npj Digit Med.* (2020) 3(41):1–4. doi: 10.1038/s41746-020-0253-3.

21. Duke Institute for Health Innovation. Health AI Partnership: an innovation and learning network for health AI software. (2021). Available from: https://dihi.org/health-ai-partnership-an-innovation-and-learning-network-to-facilitate-the-safe-effective-and-responsible-diffusion-of-health-ai-software-applied-to-health-care-delivery-settings/ (Accessed January 21, 2023).

22. Vasey B, Clifton DA, Collins GS, Denniston AK, Faes L, Geerts BF, et al. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med.* (2021) 27(2):186–7. doi: 10.1038/s41591-021-01229-5

23. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK, Keane P. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med.* (2020) 26:1364–74. doi: https://doi.org/10.1038/s41591-020-1034-x

24. Sounderajah V, Ashrafian H, Aggarwal R, Fauw JD, Denniston AK, Greaves F, et al. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI steering group. *Nat Med.* (2020) 26(6):807–8. doi: 10.1038/s41591-020-0941-1

25. Rivera SC, Liu X, Chan AW, Denniston AK, Calvert MJ, Ashrafian H, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Br Med J.* (2020) 370:m3210. doi: 10.1136/bmj.m3210

PERSPECTIVE
published: 06 July 2022
doi: 10.3389/fdgth.2022.942588

A Perspective on a Quality Management System for AI/ML-Based Clinical Decision Support in Hospital Care

Richard Bartels[1*†], *Jeroen Dudink*[2,3], *Saskia Haitjema*[4], *Daniel Oberski*[1] and *Annemarie van 't Veen*[1,5†]

[1] Digital Health, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands, [2] Department of Neonatology, Wilhelmina Children's Hospital, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands, [3] Brain Center Rudolf Magnus, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands, [4] Central Diagnostic Laboratory, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands, [5] Department of Medical Microbiology, University Medical Center Utrecht, Utrecht University, Utrecht, Netherlands

Although many artificial intelligence (AI) and machine learning (ML) based algorithms are being developed by researchers, only a small fraction has been implemented in clinical-decision support (CDS) systems for clinical care. Healthcare organizations experience significant barriers implementing AI/ML models for diagnostic, prognostic, and monitoring purposes. In this perspective, we delve into the numerous and diverse quality control measures and responsibilities that emerge when moving from AI/ML-model development in a research environment to deployment in clinical care. The Sleep-Well Baby project, a ML-based monitoring system, currently being tested at the neonatal intensive care unit of the University Medical Center Utrecht, serves as a use-case illustrating our personal learning journey in this field. We argue that, in addition to quality assurance measures taken by the manufacturer, user responsibilities should be embedded in a quality management system (QMS) that is focused on life-cycle management of AI/ML-CDS models in a medical routine care environment. Furthermore, we highlight the strong similarities between AI/ML-CDS models and *in vitro* diagnostic devices and propose to use ISO15189, the quality guideline for medical laboratories, as inspiration when building a QMS for AI/ML-CDS usage in the clinic. We finally envision a future in which healthcare institutions run or have access to a medical AI-lab that provides the necessary expertise and quality assurance for AI/ML-CDS implementation and applies a QMS that mimics the ISO15189 used in medical laboratories.

Keywords: AI, machine learning (ML), clinical decision support, implementation, quality management system, ISO15189

OPEN ACCESS

Edited by:
Sylvia Trujillo,
Independent Researcher, Portland,
OR, United States

Reviewed by:
Saifur Rahman Sabuj,
Hanbat National University,
South Korea

Correspondence:
Richard Bartels
r.t.bartels-6@umcutrecht.nl

† These authors have contributed equally to this work

Specialty section:
This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health

Received: *12 May 2022*
Accepted: *16 June 2022*
Published: *06 July 2022*

Citation:
Bartels R, Dudink J, Haitjema S, Oberski D and van 't Veen A (2022) A Perspective on a Quality Management System for AI/ML-Based Clinical Decision Support in Hospital Care. Front. Digit. Health 4:942588. doi: 10.3389/fdgth.2022.942588

INTRODUCTION

Despite the promise of new digital technologies supporting a more data-driven healthcare system, a significant gap exists between the high number of reported artificial intelligence (AI) and machine learning (ML) based algorithms in academic research and the small number of successfully implemented AI/ML-based clinical decision support (AI/ML-CDS) systems in clinical care. The valorization of AI/ML algorithms into safe and valuable AI/ML-CDS tools is considered a cumbersome process that requires broad in-depth expertise and experience in multiple domains that transcend computer-science and data analysis (1–5).

In 2017, the University Medical Center Utrecht (UMC Utrecht), one of the largest academic teaching hospitals in the Netherlands, started a hospital-wide innovation program to explore if analyses of clinical-care data could be used for AI/ML-CDS-aided personalized care. During this program, several AI/ML-CDS tools were developed in-house and some in co-creation with private parties. In this practice-oriented program, an important lesson learned was the value of a multidisciplinary approach including clinical experts, data scientists, end-users, product/service designers, software engineers, (software) security experts, ethicists, legal experts, financial/business development experts, and change management experts (6). The program evolved into the Digital Health department of the UMC Utrecht, which focuses on accelerating the implementation of digital-health technologies in clinical care for the benefit of our patients.

To support the AI/ML-CDS development process, an innovation funnel geared toward product development for use in clinical care was developed (6) and later served as a blueprint for the development of a national AI innovation tool by the Dutch Ministry of Health (7). The funnel starts with idea generation and ends with implementation in clinical care and transfer of responsibility to operational management. It is divided into seven distinctive phases with transition gates. In each phase, the relevant requirements for the specific phase are addressed including the applicable EU-laws and regulations, existing guidelines, and field standards for AI/ML development, among which are the General Data Protection Regulation (GDPR), Medical Device Regulation/*in-vitro* Diagnostic Regulation (MDR/IVDR), ISO13485 (QMS for the development of medical devices), and IEC62304 (software development lifecycle).

The GDPR, MDR/IVDR, ISO13485 and IEC62304 guidelines and standards are not explicitly developed for AI/ML-CDS tools. Efforts are undertaken to develop standards for AI/ML development (8) and numerous guidance documents exist on how to report AI/ML clinical studies (9–13). Furthermore, in a recent scoping review on guidelines and quality criteria for AI prediction models, it is acknowledged that substantial guidance is available for data preparation, model development, and model validation, while software development, impact assessment, and implementation have received less attention in scientific literature (14). Inspiration for AI/ML-lifecycle management can be gained from approaches such as CRISP-DM/ML (15–17) and contemporary software practices such as DevOps and MLOps (18, 19).

While using the national AI innovation tool as a standardized product development procedure, we have added local AI/ML-description standards, AI/ML-specific standards for version control, AI/ML audits, risk assessments, and ethical assessments. In addition, UMC Utrecht-specific templates and formats have been developed for business case analysis, stakeholder analysis, patient and customer journey analysis, data descriptions, bias risk, and so on. This way, in accordance with the core principles of MDR/IVDR, UMC Utrecht aims to direct the AI/ML-CDS development and implementation process toward a thoroughly controlled standard operating procedure (SOP) to increase the quality of the development process and its delivered products.

The Digital Health department has now progressed to implementing AI/ML-CDS tools in clinical care and this sparked a discussion on how to organize sustainable quality control of AI/ML-CDS tools within the UMC Utrecht, including roles and corresponding responsibilities of the user. ISO13485 and IEC62304 are written from the perspective of the manufacturer and are thus focused on development, implementation, and post-market surveillance procedures of the manufacturer. These guidelines appear less focused on the responsibilities of the user and the implementation of AI/ML-CDS in clinical care. Proper quality assurance requires involvement of both the manufacturer and user.

It struck us that AI/ML-CDS tools, when used as a diagnostic support system, share many similarities with clinical *in vitro* diagnostic tests used in medical laboratories. For *in vitro* devices, input material is urine, blood, or other materials, and the machine is typically a CE-marked chemical analyzer. Likewise, AI/ML-CDS input consists of data and the machine is a software system. Elaborating on this viewpoint, it is our opinion that ISO15189 for medical laboratories may serve as QMS blueprint for operating AI/ML-CDS tools in clinical practice under the MDR or IVDR. This is particularly true when used in conjunction with IEC62304. The interplay between ISO15189 and IEC62304 for software as a medical device (SaMD) under the IVDR has recently been discussed in a paper from our group (20).

In this perspective we illustrate our learnings regarding quality management of AI/ML-CDS tools through an example from our development pipeline, Sleep Well Baby (SWB). After introducing the SWB project and describing the development phase we address life-cycle management questions that arose while operationalizing SWB. When addressing these questions, we illustrate how the organizational structure of medical laboratories and ISO15189 can inspire healthcare institutes in building an effective and sustainable Quality Management System (QMS) for AI/ML usage in clinical care. Finally, in the discussion we provide an outlook how quality management of AI/ML-CDS extends to third-party AI/ML tools and settings outside healthcare institutes other than academic teaching hospitals.

SLEEP-WELL-BABY

SWB started as a grassroots project winning the best innovation price at Dutch Hacking Health 2019[1] It is an in-house developed ML model intended for monitoring real-time sleep-wake patterns in preterm neonates between 28 and 34 weeks gestational age[2] For the untrained caregiver it is almost impossible to accurately

Abbreviations: AI, artificial intelligence; CDS, clinical decision support; CRISP-DM, cross industry standard process for data mining; GDPR, general data protection regulation; IVDR, *in-vitro* diagnostic regulation; MDR, medical device regulation; ML, machine learning; NICU, neonatal intensive care unit; QMS, quality management system; SaMD, software as a medical device; SOP, standard operating procedure; SWB, sleep well baby.

[1] https://dutchhackinghealth.nl/

[2] At the time of writing the bedside implementation of SWB is still in the process of being clinically verified.

assess the sleep-wake state of preterm infants (21). The added value of real-time sleep-wake state monitoring comes from adapting elective clinical management of these preterm infants toward less disturbance during sleep periods. For a detailed discussion we refer to Sentner et al. (22).

SWB Development Phase

SWB was developed following the UMC Utrecht product innovation funnel. According to the MDR it is classified as software as a medical device class 2A, and according to the IEC62304 as category A. Being an in-house developed AI/ML-CDS, it was developed in accordance with art. 5.5 of the MDR where UMC Utrecht is both manufacturer and user. It is running at the NICU of the Wilhelmina Children's hospital (WKZ) in Utrecht and ready for use in clinical impact studies addressing how incorporating sleep-wake state information during clinical care improves patient outcomes. Development was done by a multidisciplinary development team consisting of a clinical expert, several data scientists, ML engineer, user representative and numerous experts in specific fields. During this development process, quality standards including IEC62304, ISO14971, and internal AI/ML standards were applied. Technical and clinical validation was performed by comparing predictions against a ground truth, namely sleep-wake state observations by a highly-trained and internally-calibrated team of students according to a standardized observation method (21). In **Figure 1** an overview is given of SWB development and implementation at the NICU of the WKZ. The roles and steps in the development phase are visualized on the left. Moving to the right in the figure the roles and activities in the operational phase are depicted. While transitioning to the operational phase and transferring usage and maintenance responsibilities of the SWB AI/ML-CDS tool to the clinical department, we ran into questions related to SWB life-cycle management that needed answers.

Who Is Responsible for the AI/ML-CDS Device Configuration?

The SWB configuration was developed involving multiple parties in the UMC Utrecht including the departments of Information Technology (IT), Digital Health and Clinical Physics. Each had a specific role in the development of the device configuration. In summary, the IT department provided the server and platform hosting the model, the Digital Health department data-science team provided the ML application code, and the Clinical Physics department was responsible for real-time extraction of vital parameter data from source instruments. Together with the Digital Health department they arranged the data exchange between source instrumentation, algorithm, and bedside monitor. Finally, they provide the user interface on the bedside monitor for model output. It has been decided that the Digital Health department will serve as the manufacturer and the IT and Clinical Physics departments will serve as subcontractors. The NICU serves as the user. With this division of roles, accompanying responsibilities were established and documented in SOPs and service agreements.

The questions who is responsible for which part of the configuration and who is the manufacturer are crucial in this respect. As variations exist in how AI/ML-CDS tools are configured and hosted, answers may vary per case. For example, a device can be fully developed and hosted by a third-party manufacturer, a UMC Utrecht AI/ML application can be hosted by a third-party, a third-party AI/ML application can be deployed on UMC Utrecht infrastructure, or any other variation. Agreements between parties on for instance maintenance, change management, and support during malfunction need to be addressed using a risk-based approach. ISO15189 contains several norms related to service agreements with suppliers (art. 4.6) and customers (art. 4.4).

Who Gives Clearance for the Use of SWB in Clinical Practice?

The intended use of SWB was specified by the user, the neonatologist involved. The neonatologist furthermore specified the acceptance criteria and carries responsibility for clearance of the SWB tool. Since clearance requires knowledge about both the healthcare process as well as the AI/ML model performance and its lifecycle, the clinician in charge can bear this responsibility only in consultation with a data scientist who is aware of the medical domain and can assess the device for model performance and lifecycle-management requirements. ISO15189 contains clear guidance on assigning tasks and responsibilities between employees (art 5.1).

The act of formal clearance for use needs to be repeated at specified intervals once the device is in use as part of the regular review cycle and after specific situations in which the performance of the device may be questioned, for example after observed incidents, downtime due to power failure, new releases of supportive software systems, or regular maintenance. Within the UMC Utrecht a record of AI/ML-CDS tools is kept, formal review periods are set, and standard operational qualification procedures are determined using a risk-based approach. ISO15189 contains clear norms for the introduction of equipment (art 5.3.1), reagents and disposables (art 5.3.2) and selection of examination processes (art 5.5.1) which can be extrapolated to introduction of AI/ML-CDS tools in clinical practice.

How to Ensure Safe Change Management and Revision of SWB?

As part of the development process and before implementation, an extensive risk analysis resembling a health failure mode and effect analysis on the use of the device in the care for patients within the NICU was performed. From this risk analysis agreement was reached between the stakeholders on for example forms of malfunctioning, impact of malfunctioning, and accepted downtimes. ISO15189 contains clear norms on preventive action (art 4.11).

As in-house manufacturer we applied best practices from DevOps to minimize the chance of SWB malfunction

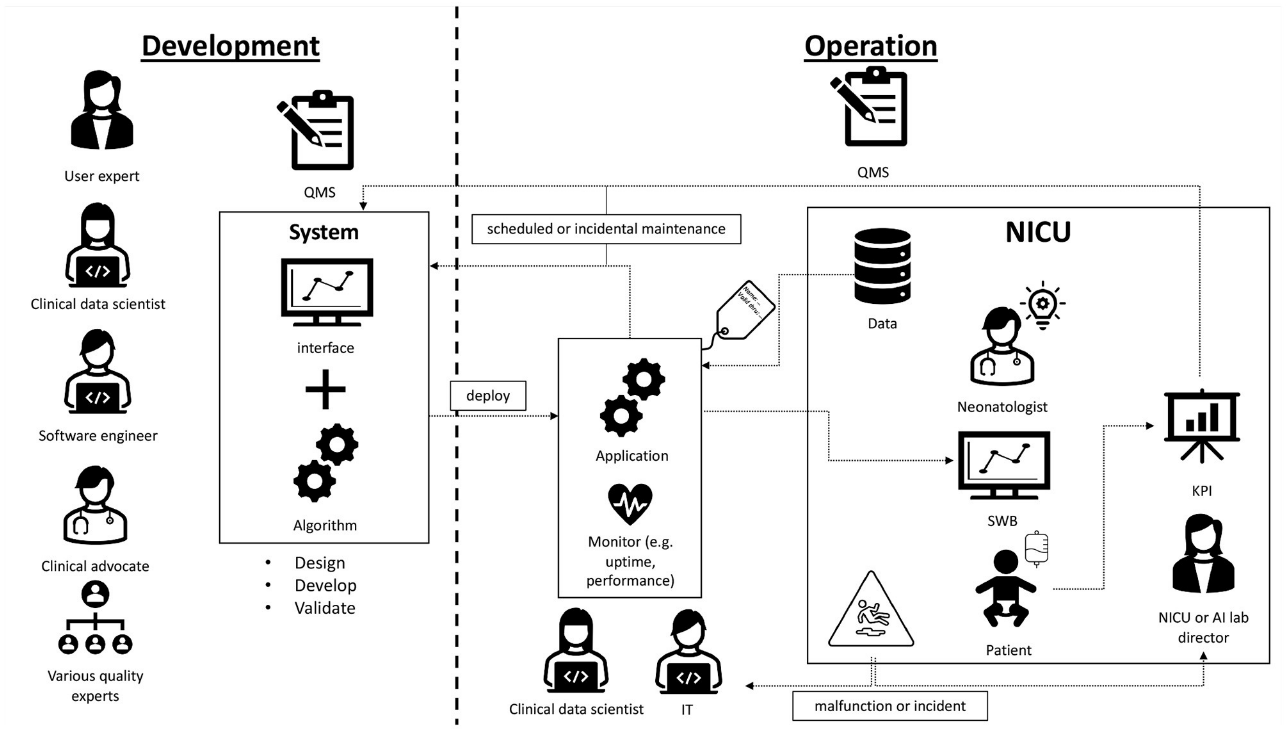


FIGURE 1 | Overview of Sleep Well Baby. Pictorial representation of how SWB was implemented on the NICU of the UMC Utrecht. The algorithm was developed by a multidisciplinary team. Currently, SWB is running bedside. It uses data from the NICU to provide sleep-wake states for preterm infants. The data scientist and software engineer remain involved for troubleshooting, monitoring and continuous maintenance. The director of the NICU is responsible for SOPs regarding AI/ML use. Governance of AI/ML-SaMDs can be done by a central AI lab with a QMS inspired by ISO15189 of the diagnostic laboratory.

and guarantee quick recovery[3]. Change management was done using *git*[4]. Data version control (*dvc*[5]) was used to ensure reproducibility and usage of the correct model in production. SWB code was extensively documented to optimize maintainability and transferability between contributors. Unit and integration tests were written for application code lowering the risk of SWB malfunctioning in clinical practice and ensuring consistency between consecutive releases. Before a change is released it first passes through mandatory review enforced by pull requests. These steps allow semi-automated and fast re-deployment of SWB. When complemented by ISO62304, ISO15189 forms a highly suitable QMS for in-house manufacturing of AI/ML-CDS tools (20).

SWB is an MDR class 2a device and carries limited patient risk. Nevertheless, appropriate procedures and responsibilities must be assigned in the SOPs of the user in case of SWB being temporarily out of service. In our role as manufacturer this implies we have an agreement with the NICU ensuring limited downtime. In practice this means that the software engineer involved in development remains involved to update SWB following the procedures specified above. This specific data science and software engineering knowledge was not transferred to the user. One can imagine that for critical devices (class 3) the user might require 24/7 support and appropriate arrangements within the organization should be established. Again, ISO15189 contains clear norms regarding the management responsibilities in providing resources to ensure quality of provided services (art 4.1.2).

What if Model Performance Starts Degrading?

Predictive models can degrade over time due to their dependence on input data from potentially changing environments or self-induced feedback loops. Consequently, AI/ML models require monitoring of model performance. During the AI/ML risk analysis, the question was asked: what are the chances of SWB performance degrading? Which process mitigation measures can be applied? And what to do in case of degradation?

[3] The DevOps movement is the current paradigm in software development, combining development (Dev) and operations (Ops) teams for increased efficiency throughout the software lifecycle (18). IEC62304 is sometimes believed to hamper the use of contemporary software development practices such as DevOps (26). However, we believe that agile DevOps change management practices can be successfully combined with the MDR and IVDR, which prescribe the use of generally acknowledged state-of-the-art technologies. Moreover, we are of the opinion that activities prescribed by IEC62304 and the quality control measures they enforce can be successfully incorporated within the DevOps philosophy.

[4] https://git-scm.com

[5] https://dvc.org

SWB is a locked algorithm[6]. Since it only depends on vital parameters, major performance degradation was considered unlikely in the risk analysis. Nevertheless, a change in hardware collecting vital parameters or a changing patient population could result in model drift. The user should be aware of this risk and should be capable to identify it on occurrence. The manufacturer should inform users of this risk in general and specifically in relation to the context in which the AI/ML-CDS tool is used. Building on best-practices from the MLOps movement a monitoring dashboard was designed for SWB, tracking the fraction of valid requests to the model service and tracking distributions of predicted sleep-wake states over time. These distributions serve as a proxy for model performance in absence of a direct accuracy measurement (no other sleep-wake state measurements are performed with regular intervals). Monitoring model performance is, contrary to application performance, not a requirement of the IEC62304, but its relevance is acknowledged (23). In case of degrading model performance, a decision should be made by the user to either (temporarily) terminate the application and/or to re-calibrate and re-validate SWB.

SWB monitoring and re-calibration of the model is done by the Digital Health department since they have the appropriate procedures and competencies. Again, monitoring and re-calibration requires the expertise of data scientists. Furthermore, since UMC Utrecht is the manufacturer and user, we have access to the required data to perform monitoring. However, for most manufacturers this will not necessarily be the case. In this situation the manufacturer could make available tooling for monitoring and re-calibration, or the user should set up monitoring procedures themselves. **Figure 1** on the right depicts the continuous involvement of the data scientist in monitoring the application.

Who Provides a Helpdesk for Users?

Sections How to ensure safe change management and revision of SWB? and What if model performance starts degrading? discussed malfunction and model degradation. This raises the question, what if a user experiences a malfunction? Or what if an incident involving SWB occurs? The user is responsible for having appropriate incident management, in addition to the post market surveillance responsibilities of the manufacturer. Feedback of incidents affecting patient care is already covered by existing NICU procedures. For malfunctions not directly affecting the patient a SWB helpdesk was created. Here reports can be filed and will be handled by the appropriate experts, such as described in the previous section and illustrated in **Figure 1**.

How Are Users Trained?

A prospective risk analysis performed by the user revealed the risk of SWB being incorrectly used due to imperfect model performance and raised the question: how can this be prevented? SWB is a sleep-wake monitoring system intended primarily for nurses to plan elective care (e.g., changing diapers). It differs from other monitors-such as heart rate-in that it is not based on direct physiological measurement but instead makes a prediction with imperfect precision. In addition, it was developed for a particular population of preterm infants, i.e., inclusion criteria. Nurses and neonatologists should be aware of these limitations such that they can use the device appropriately. The NICU should ensure appropriate SOPs for SWB, including procedures on disregarding SWB advice. Meanwhile, the manufacturer should provide user instructions and guidance documentation specifying amongst other things the intended use, mode of operation, intended patient population and limitations in terms of sensitivity and specificity. This is similar to instructions included with medication or *in vitro* devices. User-employed specialists or the manufacturer should provide training and guidance to end-users when required. In the medical lab it is customary to organize a training by the manufacturer with the introduction of a new analyzer. After the introduction of the analyzer new employees are trained internally by internal employees who are competent in operating the analyzer. ISO15189 provides clear norms on training programs for employees (5.1.5) and monitoring and assessing competences of employees (5.1.6) which can be extrapolated to AI/ML-CDS usage.

DISCUSSION AND CONCLUSION

In the context of SWB, we discussed a selection of quality aspects and responsibilities that surface when operating AI/ML-CDS in clinical practice. We showed how ISO15189 can be a source of inspiration for a healthcare institute its QMS for operating and in-house manufacturing of AI/ML-CDS tools. UMC Utrecht is learning-by-doing, SWB is only a first example and the effort of implementing quality measures to ensure safe use of AI/ML-CDS tools in clinical practice is still in progress. Moreover, the AI/ML field itself is still maturing and quickly evolving.

SWB is an in-house developed ML algorithm where UMC Utrecht is both manufacturer and user. The extrapolation to AI/ML purchased from a third-party is relatively straightforward. Manufacturers should adhere to a QMS for production such as ISO13485. Users of third-party devices are accountable for responsible use of AI/ML-CDS, their QMS should include processes for selection, clearance and performance verification, appropriate SOPs, and service agreements with the manufacturer relating to monitoring and change management. ISO15189 could provide inspiration for this. It is of great importance that the user has the appropriate expertise to audit (24) and validate AI/ML-CDS tools or else a situation can arise where underperforming and potentially harmful use of AI/ML in clinical practice is not being identified (25). In case departments of a healthcare institution are unable to provide this expertise themselves, it could be bundled in a centralized AI laboratory.

[6] There is a distinction between locked and adaptive algorithms. Locked algorithms are static functions. SWB is a static classifier, given the same input data it will always return the same result. A locked algorithm can be re-calibrated or updated manually in an *ad-hoc* fashion, for example when introduced to a new ward or when a larger dataset becomes available. On the contrary, adaptive algorithms are continuously updated through a (semi-)automatic process. In theory, such algorithms can adapt automatically to a changing environment to prevent model drift.

Our recommendations hold true for larger healthcare institutions such as academic teaching hospitals who can build the necessary resources and competences needed for safe operation of AI/ML-CDS tools. For smaller entities, such as a single general practitioner, this effort seems unfeasible. In this situation, complete dependence on the manufacturer is imaginable, making it difficult to establish truly safe performance. Again, inspiration can be found in the regional services of medical laboratories that very often provide access to competences and resources for safe application of diagnostics. Regional AI labs could provide services for the development, acquisition, and quality control of AI/ML for smaller healthcare institutes including general practitioners. Like medical laboratories they could educate and assist healthcare professionals in the selection and safe use of AI/ML.

Complying with an extensive user QMS is time-intensive, expensive, and might appear to hamper innovation. However, just like *in vitro* devices, an appropriate QMS is a necessity for safe AI/ML use within healthcare settings. In spirit with the MDR/IVDR it is quality first. Moreover, so far AI/ML has not yet lived up to its promise to revolutionize healthcare. Although we believe it has the potential to do so, we do not envision a disruptive change in which dozens of AI/ML-CDS systems will independently enter every department in the coming years. Instead, it will more likely be a regulated introduction similar in pace to the way new *in vitro* devices or medication are introduced. We strongly believe an appropriate QMS will not only guarantee safe use, but also helps accelerate implementation. The lessons learned and identified quality criteria in this perspective illustrate that ISO15189 can serve as an inspiration and provide a starting point for organizations building their own data-driven capacity to improve patient care.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author.

AUTHOR CONTRIBUTIONS

RB, DO, and AV initiated this perspective. JD is the neonatologist in charge of the SWB project. RB and AV drafted the manuscript SH provided extensive feedback on the manuscript. All authors took part in discussion, revision of the manuscript, contributed to the article, and approved the submitted version.

REFERENCES

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. *Nat Med*. (2019) 25:44–56. doi: 10.1038/s41591-018-0300-7
2. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med*. (2019) 25:30–6. doi: 10.1038/s41591-018-0307-0
3. Matheny ME, Whicher D, Thadaney Israni S. Artificial intelligence in health care. *JAMA*. (2020) 323:509. doi: 10.1001/jama.2019.21579
4. Wolff J, Pauling J, Keck A, Baumbach J. Success factors of artificial intelligence implementation in healthcare. *Front Digit Heal*. (2021) 3:1–11. doi: 10.3389/fdgth.2021.594971
5. van de Sande D, van Genderen ME, Huiskens J, Gommers D, van Bommel J. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. *Intensive Care Med*. (2021) 47:750–60. doi: 10.1007/s00134-021-06446-7
6. Haitjema S, Prescott TR, van Solinge WW. The applied data analytics in medicine program: lessons learned from four years' experience with personalizing health care in an academic teaching hospital. *JMIR Form Res*. (2022) 6:1–5. doi: 10.2196/29333
7. Publicatie. Data voor gezondheid. *Innovation Funnel for Valuable AI in Healthcare*. (2021). Available online at: https://www.datavoorgezondheid.nl/documenten/publicaties/2021/07/15/innovation-funnel-for-valuable-ai-in-healthcare (accessed March 23, 2022).
8. ISO - ISO/IEC CD 5338. *Information technology—Artificial intelligence—AI system Life Cycle Processes*. Available online at: https://www.iso.org/standard/81118.html (accessed April 21, 2022).
9. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. (2020) 26:1364–74. doi: 10.1038/s41591-020-1034-x
10. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ, Darzi A, et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med*. (2020) 26:1351–63. doi: 10.1038/s41591-020-1037-7
11. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digit Med*. (2020) 3:1–10. doi: 10.1038/s41746-020-0221-y
12. Collins GS, Dhiman P, Andaur Navarro CL, Ma J, Hooft L, Reitsma JB, et al. Protocol for development of a reporting guideline (TRIPOD-AI) and risk of bias tool (PROBAST-AI) for diagnostic and prognostic prediction model studies based on artificial intelligence. *BMJ Open*. (2021) 11:1–7. doi: 10.1136/bmjopen-2020-048008
13. Shelmerdine SC, Arthurs OJ, Denniston A, Sebire NJ. Review of study reporting guidelines for clinical studies using artificial intelligence in healthcare. *BMJ Heal Care Informatics*. (2021) 28:1–10. doi: 10.1136/bmjhci-2021-100385
14. de Hond AAH, Leeuwenberg AM, Hooft L, Kant IMJ, Nijman SWJ, van Os HJA, et al. Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review. *NPJ Digit Med*. (2022) 5:2. doi: 10.1038/s41746-021-00549-7
15. Shearer C. The CRISP-DM model: the new blueprint for data mining. *J Data Warehous*. (2000) 5:13–22.
16. Studer S, Bui TB, Drescher C, Hanuschkin A, Winkler L, Peters S, et al. Towards CRISP-ML(Q): a machine learning process model with quality assurance methodology. *Mach Learn Knowl Extr*. (2021) 3:392–413. doi: 10.3390/make3020020
17. Kolyshkina I, Simoff S. Interpretability of machine learning solutions in public healthcare: the CRISP-ML approach. *Front Big Data*. (2021) 4:660206. doi: 10.3389/fdata.2021.660206
18. Forsgren N, Humble J, Kim G. *Accelerate: The science behind devops : building and scaling high performing technology organizations*. IT Revolution (2018). Available online at: https://books.google.nl/books?id=85XHAQAACAAJ
19. *MLOps: Continuous Delivery and Automation Pipelines in Machine Learning. Google Cloud*. Available online at: https://cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning (accessed January 27, 2022).
20. van Deutekom HWM, Haitjema S. Recommendations for IVDR compliant in-house software development in clinical practice: a how-to paper with three use cases. *Clin Chem Lab Med*. (2022) 60:982-8.doi: 10.1515/cclm-2022-0278

21. de Groot ER, Bik A, Sam C, Wang X, Shellhaas RA, Austin T, et al. Creating an optimal observational sleep stage classification system for very and extremely preterm infants. *Sleep Med.* (2022) 90:167–75. doi: 10.1016/j.sleep.2022.01.020
22. Sentner T, Wang X, de Groot ER, van Schaijk L, Tataranno ML, Vijlbrief DC, et al. *The Sleep Well Baby Project: An Automated Real-Time Sleep-Wake State Prediction Algorithm in Preterm Infants.* Submitted (2022).
23. Data voor gezondheid. Publicatie. *Guideline for High-Quality Diagnostic and Prognostic Applications of AI in Healthcare*. Available online at: https://www.datavoorgezondheid.nl/documenten/publicaties/2021/12/17/guideline-for-high-quality-diagnostic-and-prognostic-applications-of-ai-in-healthcare (accessed March 23, 2022).
24. Oala L, Murchison AG, Balachandran P, Choudhary S, Fehr J, Leite AW, et al. Machine learning for health: algorithm auditing & quality control. *J Med Syst*. (2021) 45:105. doi: 10.1007/s10916-021-01783-y
25. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med.* (2021) 181:1065–70. doi: 10.1001/jamainternmed.2021.2626
26. McHugh M, McCaffery F. Adopting agile practices when developing medical device software. *Comput Eng Inf Technol.* (2015) 4:2. doi: 10.4172/2324-9307.1000131

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

NOMENCLATURE

International Organization for Standardization (ISO) Norms

IEC 62304:2006+A1:2015
Medical device software–Software life cycle processes
ISO 13485:2016
Medical devices–Quality management systems–requirements for regulatory purposes
ISO 15189:2012
Medical laboratories–Requirements for quality and competence.

TYPE Original Research
PUBLISHED 04 August 2022
DOI 10.3389/fdgth.2022.932411

An integration engineering framework for machine learning in healthcare

Azadeh Assadi[1,2]*, Peter C. Laussen[3,4], Andrew J. Goodwin[1,5], Sebastian Goodfellow[6], William Dixon[1], Robert W. Greer[1], Anusha Jegatheeswaran[7], Devin Singh[8,9], Melissa McCradden[10,11,12], Sara N. Gallant[1], Anna Goldenberg[12,13,14,15], Danny Eytan[1,16,17†] and Mjaye L. Mazwi[1,3,8†]

[1]Department of Critical Care Medicine, Hospital for Sick Children, Toronto, ON, Canada, [2]Institute of Biomaterials and Biomedical Engineering, Department of Engineering and Applied Sciences, University of Toronto, Toronto, ON, Canada, [3]Institute of Medical Sciences, University of Toronto, Toronto, ON, Canada, [4]Executive Vice President for Health Affairs, Boston Children's Hospital, Boston, MA, United States, [5]School of Biomedical Engineering, University of Sydney, Sydney, NSW, Australia, [6]Department of Civil and Mineral Engineering, Faculty of Applied Science and Engineering, University of Toronto, Toronto, ON, Canada, [7]Department of Surgery, Division of Paediatric Cardiac Surgery, Hospital for Sick Children, Toronto, ON, Canada, [8]Translational Medicine, Peter Gilgan Centre for Research & Learning, Toronto, ON, Canada, [9]Department of Emergency Medicine, The Hospital for Sick Children, Toronto, ON, Canada, [10]Department of Bioethics, The Hospital for Sick Children, Toronto, ON, Canada, [11]Division of Clinical and Public Health, Dalla Lana School of Public Health, Toronto, ON, Canada, [12]Genetics & Genome Biology, Peter Gilgan Centre for Research & Learning, Toronto, ON, Canada, [13]Department of Computer Science, University of Toronto, Toronto, ON, Canada, [14]Vector institute for Artificial Intelligence, University of Toronto, Toronto, ON, Canada, [15]CIFAR, Toronto, ON, Canada, [16]Department of Medicine, Technion, Haifa, Israel, [17]Department of Pediatric Critical Care, Rambam Medical Center, Haifa, Israel

OPEN ACCESS

EDITED BY
Mark Sendak,
Other, United States

REVIEWED BY
Carly Bobak,
Dartmouth College, United States
Thankam Paul Thyvalikakath,
Indiana University, United States

*CORRESPONDENCE
Azadeh Assadi
az.assadi@mail.utoronto.ca

†These authors share senior authorship

SPECIALTY SECTION
This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health

RECEIVED 29 April 2022
ACCEPTED 18 July 2022
PUBLISHED 04 August 2022

CITATION
Assadi A, Laussen PC, Goodwin AJ, Goodfellow S, Dixon W, Greer RW, Jegatheeswaran A, Singh D, McCradden M, Gallant SN, Goldenberg A, Eytan D and Mazwi ML (2022) An integration engineering framework for machine learning in healthcare. Front. Digit. Health 4:932411. doi: 10.3389/fdgth.2022.932411

Background and Objectives: Machine Learning offers opportunities to improve patient outcomes, team performance, and reduce healthcare costs. Yet only a small fraction of all Machine Learning models for health care have been successfully integrated into the clinical space. There are no current guidelines for clinical model integration, leading to waste, unnecessary costs, patient harm, and decreases in efficiency when improperly implemented. Systems engineering is widely used in industry to achieve an integrated *system of systems* through an interprofessional collaborative approach to system design, development, and integration. We propose a framework based on systems engineering to guide the development and integration of Machine Learning models in healthcare.

Methods: Applied systems engineering, software engineering and health care Machine Learning software development practices were reviewed and critically appraised to establish an understanding of limitations and challenges within these domains. Principles of systems engineering were used to develop solutions to address the identified problems. The framework was then harmonized with the Machine Learning software development process to create a systems engineering-based Machine Learning software development approach in the healthcare domain.

Results: We present an integration framework for healthcare Artificial Intelligence that considers the entirety of this *system of systems*. Our proposed framework utilizes a combined software and integration engineering approach and consists of four phases: (1) Inception, (2) Preparation, (3) Development, and (4) Integration. During each phase, we present specific elements for consideration in each of the three domains of integration: *The Human, The Technical System,* and *The Environment.* There are also elements that are considered in the interactions between these domains.

Conclusion: Clinical models are technical systems that need to be integrated into the existing *system of systems* in health care. A systems engineering approach to integration ensures appropriate elements are considered at each stage of model design to facilitate model integration. Our proposed framework is based on principles of systems engineering and can serve as a guide for model development, increasing the likelihood of successful Machine Learning translation and integration.

KEYWORDS

Integration engineering, artificial intelligence, machine learning, digital health, system of systems (SoS), human factors engineering (HFE), healthcare (MeSH)

Glossary of key terms used in this manuscript

Artificial Intelligence (AI): A field that combines computer science and robust datasets to solve problems. These systems can be said to think like humans, act like humans, think rationally, or act rationally (1).

Food and Drug Administration (FDA): Regulatory body of the government responsible for maintaining health and safety of humans and animals within the United States through regulating food, drug, and technology (2).

Human Factors Engineering (HFE): A field of engineering focusing on the design of technology tailored to people as well as sociotechnical integration involving large, complex systems such as healthcare (3, 4).

ISO/IEC/IEEE: Global product, technological, procedural, and engineering standards set by globally recognized ISO/IEC/IEEE organizations (5, 6).

Machine Learning (ML): A branch of artificial intelligence and computer science that describes the ability of an algorithm to "learn" by finding patterns in large datasets (7).

Software Development Lifecycles (SDLC): The various software development frameworks that are used to structure software development (8).

System of Systems (SOS): A system that is composed of other systems and its "elements are managerially and/or operationally independent" (9).

Introduction

Artificial Intelligence offers transformational opportunities in medicine, but this potential remains limited by a translation gap (10, 11). There are a variety of drivers that contribute to this gap. First, restrictions in data sharing limit training, validation and improvement of models (12). Second, lack of data and model transparency limit clinicians' ability to interpret the model and evaluate it for relevance, accuracy and bias impacting their trust in the model and thereby limiting utilization (12–14). Third, the absence of established model verification processes impose further challenges (12–14). Financial constraints, limited physician training in the field and rules and regulations that often lag technological advances further affect model integration (12–14). While the Food and Drug Administration has proposed guidelines to regulate some clinical models (15, 16), there are no current guidelines for clinical model integration. The term "model integration" is a more appropriate term than "implementation" as it recognizes that Artificial Intelligence models need to be compatible with the complex sociotechnical environments that characterize healthcare. Integration is defined as "an act or instance of combining into an integral whole" and refers to combining several implemented elements to form a fully realized system that enables interoperability between the various elements of the system (9). Improper integration of new systems may lead to additional costs, patient harm, damage to other systems, and decrease in efficiency (9). To address this translation gap, we present a systems engineering framework to guide the development of models with explicit consideration of elements that are crucial for successful model integration in healthcare.

Foundations

Systems engineering

Systems engineering is an interdisciplinary approach to system design to ensure its interactive elements are organized to achieve the purpose of the system (17, 18). The term dates to the early 1940's and Bell Telephone Laboratories where it was used during World War II (19). The need for systems engineering came from the discovery that satisfactory components did not necessarily combine to produce a satisfactory system (20). This was particularly a problem for industries which produced complex systems at an early date, such as communications and aircraft industries (20). Systems engineering forms the foundations of ISO/IEC/IEEE and is currently applied in a wide variety of industries from manufacturing to engineering and aerospace (18). A system is defined as "*an aggregation of elements organized in some structure to accomplish system goals and objectives, is usually composed of* ***humans and machines*** *and has a definable structure and organization with external boundaries that separate it from elements outside the system* (4)." A Machine Learning model should really be considered as a system composed of the model and data sources, its users, and its context. At higher degrees of abstraction, a system can be composed of other systems to create a system of systems (SOS), defined as a system whose "*elements are managerially and/or operationally independent*" (9). The healthcare system is a SOS (**Figure 1**) that results from the integration of the medical system, the regulatory, legal, and ethical systems, the financial system, the hospital system, the electronic health records, and many others. Therefore, for applied Machine Learning to become part of this existing SOS, it must be effectively integrated. Integration is the key to viability of any SOS (21) and achieving it requires effective collaboration between these inter-operable systems (9). Integration, therefore, is a process that is analyzed, planned, designed, developed, executed, managed, and monitored throughout the system's entire lifecycle and not just a distinct phase at the end (9).

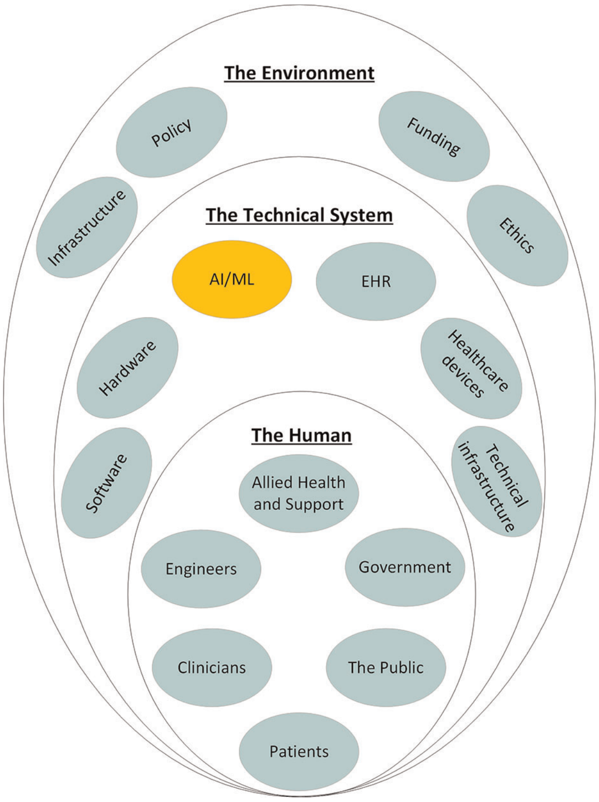


FIGURE 1
System of systems of healthcare and how applied machine learning should integrate into these existing systems. Each of the elements shown here influence every other element in an interconnected network. Electronic Health Record (EHR); Artificial Intelligence/ Machine Learning (AI/ML).

Integration should consist of integrating the technical aspects of the system as well the human-system aspects. Technical integration of systems ensures that the various technical aspects of the system can work together to achieve the common objective (9). In the context of applied Machine Learning, this requires the interoperability of the models with the existing hardware and software infrastructure. To facilitate the technical integration of software, modular programming styles as well as Software Development Life Cycle (SDLC) frameworks have mechanisms that require developers to understand the existing systems and evaluate how the new system fits within this existing system as they progress through iterative development cycles. This is particularly important in dynamic industries like healthcare. The human-system integration aspect, also known as the sociotechnical integration, refers to the integration of a system within the social construct of the environment (9). Social demands as well as the societal or cultural values can play a major role in determining optimal performance of the entire system (22). Sociotechnical integration can be achieved through Human Factors Engineering (HFE), which is a field of engineering focusing on the design of technology tailored to people as well as sociotechnical integration involving large, complex systems such as healthcare (3, 4). The Systems Engineering Initiative for Patient Safety model provides a framework for integrating HFE in to healthcare to improve patient care quality and safety (23). Currently in its third iteration, this framework strongly advocates for a human centered design approach to engineering various aspects of the health care system such that the needs of patient and the people who care for these patients are put at the center of the design process (23).

Within systems engineering, the domains of integration are *The Technical System, The Human,* and *The Environment* and the interactions between them which have been previously described in literature (24) and summarized in **Table 1**. While this framework for integration has been used in the industry, it has not been applied to integrating Machine Learning in healthcare.

Systems engineering in software development

Systems engineering principles are applied in software development. The SDLC defines the development process of software (8). It also relates to the architecture of the software and facilitates an understanding of the required resources for the software (8). The use of agile software development

TABLE 1 Domains of integration and the interaction between them.

Domains of Integration	Definition
The Technical System	“An aggregation of elements organized in some structure to accomplish system goals and objectives, is usually composed of humans and machines and has a definable structure and organization with external boundaries that separate it from elements outside the system” (4)
Human	“An individual, a group of individuals, or organizations which have connections to the system in the form of owners, users, operators, managers, service providers, supplies, producers, or other stakeholders, who directly or indirectly have an interest in the system.” (9)
Environment	“All the relevant parameters that can influence or be influenced by the system in any lifecycle phase.” (9)
System-Environment interaction	A physical interaction occurs through technical interfaces while a non-physical interaction can occur through laws, regulations, policy, market demands and political interests, which may influence or be influenced by the system (9).
Human-System interaction	The physical, logical, or emotional relationship between the human and the system that can be influenced by or influence the system. HFE largely aims to optimize this interaction (9, 25).
Human-Environment interaction	Relationship between the human and the internal and external workplace or system environment. Some examples include organizational attributes that may affect decision-making processes of humans, circumstances that may cause deviation from standard operating procedures, impact of noise, temperature, illness, fatigue, interpersonal relationships, etc. can also influence the system or be influenced by it (9, 25). HFE can also be used here to optimize some of these challenges.

HFE, Human Factors Engineering.

techniques has facilitated software implementation but is seldom used by healthcare Machine Learning model developers. This approach also fails to recognize some of the unique implications of applied Machine Learning in practice. The purpose of SDLC has evolved over its 60-year history from ensuring an understanding of what needs to be done to focusing on structured development methods, to focusing on product delivery (26). To achieve this, there needs to be a balance between the structured and agile SDLC frameworks (26). The evolution of these different life cycles has been in response to the increasing complexity of software, the systems for which new software are being designed, advancements in hardware, and the widespread use of software in society. Despite successful use of software globally, a strong emphasis on time to market has led to the incomplete application of many well-known SDLC recommendations, particularly those of requirement gathering, planning, specifications, architecture, design, and documentation (27). There are also inherent limitations to each of the SDLC models that can further contribute to poor software design (**Table 2**). In addition to cost, technical concerns, need for workflow alterations, privacy concerns, perceived lack of usefulness, productivity loss, and usability issues have contributed to the very slow uptake of healthcare software such as electronic health record systems in United States (30).

Challenges of machine learning and existing software development life cycle

There are multiple frameworks proposed for Machine Learning development that essentially focus on *context understanding, data curation, data modeling,* and *production and monitoring* (31–34). Yet, despite relatively fast and cheap development and deployment of Machine Learning models, they have been difficult and expensive to maintain and integrate (35). This is in part due to existing challenges with medical software and in part due to some unique Machine Learning issues (35). **Table 3** illustrates some of the Machine Learning challenges that impact its development.

Challenges with machine learning model integration into healthcare

Developing Machine Learning models for healthcare imposes unique challenges that can impact successful clinical integration. Some challenges relate to the social complexity of medicine and others to the safety critical nature of medical systems. **Table 4** summarizes some of the described challenges and gaps in clinical Artificial Intelligence models and the health care environment that limit their use. For example, health care data can be very noisy and as such, often subject to data preprocessing. This preprocessing may result in training and testing data that may not be representative of the “real world” data that the model will experience in practice (**Table 3**, M14; **Table 4**, C7). Variations in institutional Electronic Health Record and the information in the structure of this type of data also challenge model performance across institutions (45).

Methods

A modified narrative review method was used to understand (a) the challenges and gaps in integrating Machine Learning models in health care, (b) challenges associated with Machine learning models and current SDLC, and (c) principles of

TABLE 2 Overview of current software development life cycle models and their limitations (8, 28, 29).

SDLC	Overview	Limitations
Classical Waterfall model	- Series of processes in succession without gap - Foresees defect or fault - Requires proper planning and well-articulated documentation - "Characterize before the design" - Used in safety critical systems where phases and processes are inter-dependent and there is a high need for assurance with no tolerance for mistakes	- Not flexible - Prototypes made late in the overall process - Product delivery often delayed - High risk and uncertainty - Not suitable for complex and object-oriented projects - Not suitable for long and ongoing projects - Not suitable for existing systems
Iterative Waterfall model	- Use iterations to prototype and refine the project's requirements before proceeding with the waterfall model for the rest of the development process	- Iteration is possible but predisposed to errors and costly - Not suitable for long term projects - Difficult to gather requirements - Changes in previous stages can cause big issues in subsequent stages
Prototyping model	- Leverage the use of prototypes to clarify and refine requirements - May use prototype to iteratively build the finished project or simply use it as a demonstration of what is being proposed as a solution	- Requires system modifications after implementation - Can increase complexity of the system - Leads to incomplete applications
Evolutionary model	- Requirements change over time and the initial design evolves with user interaction and input as well as with new requirements	- Planning and design phase are incomplete - Not suitable for incremental building - Costly
Spiral model	- Combination of top-down and bottom-up constructs - Can be used with other models - Breaks a project into smaller segments so simplify development and evaluation - For systems when cost and risk assessment are key - Also, when users are uncertain about their needs	- Costly - Requires high expertise for risk analysis - Risk analysis central to project success - Not suitable for small projects
V-model	- Focus on validation and verification—the product from each phase is checked and approved before moving on to the next phase	- Very rigid - Prototypes are available late in the development phase - Changes require lots of documentation
RAD model	- Rapid, iterative design of small parts of the project to put into test and ensure project on track and meeting requirements before pursuing the next iteration - Agile software development falls in this category	- Depends on strong member performance to identify requirements - Only suitable for modular systems - Requires very skilled developers with good modeling skills - Costly

SCLD, Software Development Life Cycle; RAD, Rapid Application Development.

integration engineering from a systems engineering and human factors engineering perspective. A narrative review method is the searching of the literature with a specific goal in mind where manuscripts are hand-selected for inclusion based on the research questions (46). The ACM-DL, PubMed, and IEEE Xplore databases were used for this narrative review with the following search terms: ("All Metadata":"machine learning" OR "All Metadata":"artificial intelligence" OR "All Metadata":"algorithm" OR "All Metadata":"model"), ("All Metadata":"healthcare" OR "All Metadata":"medicine" OR "All Metadata":"clinical" OR "All Metadata":"health care" OR "All Metadata":"health"), ("All Metadata":framework), ("All Metadata":development), and ("All Metadata":"software development lifecycle"). The search was restricted to publications in the last 5 years and original peer-reviewed research and reviews. Duplicate results were removed based on manuscript title, and relevant manuscripts were selected based on abstracts. The selected manuscripts were then reviewed, summarized, and synthesized to outline (a) challenges and gaps in integrating Machine Learning Models in health care, (b) challenges associated with Machine learning models and current SDLC, and (c) principles of integration engineering from a systems engineering and human factors engineering perspective.

To address some of the challenges associated with Machine Learning models and current SDLC as well as Machine

TABLE 3 Summary of challenges associated with machine learning life cycle (35).

Aspects of software Development	Features and challenges with Machine Learning development
Software Requirements	- Uncertain requirements (conceptual description of the goal after applying Machine Learning systems; different data and different application context would lead to different requirements) [M1] - Quantitative measures such as accuracy define requirements with little regard to functional requirements (the exact desired quantitative measures (e.g., accuracy) are not always known) [M2] - Requirement validation requires a larger number of preliminary experiments, ideally with real data [M3] - Requirement must consider the predictable degradation in performance of Machine Learning systems (must be degradation-sensitive and adapt to degradation through ongoing training or re-training) [M4]
Software Design	- Insufficient emphasis on the coupling of components (e.g., quality of data processing and performance of Machine Learning models) [M5] - Flexible detailed design with need for multiple, iterative experimentation to develop an effective model [M6]
Software Construction and Tools	- Bulk of coding is focused on developing an effective Machine Learning model [M7] - Debugging focused on improving model performance (need real data and often delayed until last stages) [M8] - Debugging can take a very long time based on data size and complexity of a model [M9] - Bugs can be hidden in the data [M10]
Software Testing and Quality	- Hard to reproduce test results because of sources of randomness [M11] - Testing output are often a range or probability based rather than a single value [M12] - Quality of testing is highly dependent on the quality of the test case and testing dataset [M13] - Good testing results cannot guarantee performance in production or generalizability (highly dependent on similarities of training/testing datasets and the real-world data) [M14]
Software Maintenance and Configuration Management	- Expect performance degradation [M15] - Require configuration management to keep track of varying models and associated tradeoffs, algorithm choice, architecture, data, hyperparameters, etc. [M16]
Software engineering Process and Management	- Overestimation of what Machine Learning can do leading to mismatch of expectation and reality [M17] - Limited incorporation of domain expertise into the engineering and management process [M18] - Sustained performance requires ongoing monitoring and planned evaluation to determine timing to retrain and to rectify mistakes and unexpected consequences [M19] - No standard guidance for the management of Machine Learning development [M20]

ML, Machine Learning.

Learning models for the health care environment, we adapted existing software development guidelines as well as principles of systems engineering to develop our framework described below. The framework was iteratively designed through synthesis of the literature with expert input from the research team in domains of human factors engineering, Machine Learning, medicine, and software development. Consensus was achieved on the final iteration of the framework which was organized in keeping with steps in Machine Learning model development. In the supplementary material, we apply this framework to the design of an arrhythmia detection model intended for clinical integration.

Results

We present a generalizable framework (Figure 2) that identifies four phases in the development of Artificial Intelligence models in healthcare: (1) Inception, (2) Preparation, (3) Development, and (4) Integration. Each phase incorporates considerations from the key domains of integration and systems engineering as well as the interaction between them for an integrated SOS as we show below. Outcomes from each phase while informing the phase that follows, also provides feedback to previous phases, particularly when there are new findings in a phase that were not previously considered. The challenges outlined in Table 3 that are addressed with our proposed model are indicated in [M#] format while challenges from Table 4 are indicated in [C#] as the features and steps are described. We have applied this framework to an arrhythmia detection model and have implemented this as a best practice at the Hospital for Sick Children. This practical application is demonstrated in the supplementary material. It is important to note, that the successful integration of a system into an existing system also requires ongoing maintenance and refinement which includes detection of performance degradation, changing workflows and policies, changing hardware and data acquisition, and

TABLE 4 Challenges with machine learning models in healthcare.

Aspects of Machine Learning-models and the healthcare environment	Gaps or Challenges
Context	- Need to thoroughly understand the clinical data being used for model development (13, 36) [C1] - Need models with impactful clinical utility (13) [C2] - Need models that fit within the environment they are intended for (13, 37–41) [C3]
Data	- Need access and availability to well labeled, high quality, large datasets (13, 14, 39) [C4] - Need consistency in data collection techniques (13) [C5] - Need to acknowledge and minimize inaccurate or incomplete data (13, 41) [C6] - Need to ensure that model training/test data is representative of what the model will experience during operation; consider pre-processing of data and its effect [C7] - Need to identify, remove, and account for biased data (13, 14, 40, 41) [C8] - Need to account for data shifts and their effect on model performance [C9]
Model validation and performance	- Need to conduct and develop clinical validation studies (11, 13, 14, 37) [C10] - Need to conduct clinical impact/outcome studies as Machine Learning metrics (accuracy, precision, etc.) often do not map directly to clinical performance indicators (14, 37) [C11] - Need model transparency (11, 39, 41) [C12]
Ethics and Regulation	- Need regulation and safe use guidelines (14, 39, 42) [C13] - Need privacy and cybersecurity regulations (39–41, 43) [C14] - Need to screen for algorithmic biases (11) [C15]
Financial issues	- Need adequate resources (hardware, expertise, software, etc. all in high demand, limited, and expensive) to develop and integrate models (39) [C16]
Knowledge gap	- Need users to have sufficient knowledge to interpret model output or compare different models (11, 39, 41, 44) [C17]

ML, Machine Learning.

changing knowledge and familiarity with Artificial Intelligence [M15, M19, M20]. This maintenance phase is not included in our integration framework as this follows successful initial model integration which is our focus.

Inception

Inception refers to the very first phase of model development during which an appropriate use case and modelling approach are identified. During this phase, specific considerations for the *Technical System* include clear problem definition and clinical applications [M1, C1 & C2], strategies, and techniques to address the problem [M1], the kinds of data the model would need and whether the data is available or can be made available [M5, M13 & M14], sourcing the data, and whether the data considered for training is similar to what would be used in the intended environment [M5, M13, M14, C1-4, C7]. Close collaboration between clinicians (the domain experts and future users), model developers, and data scientists is essential. The optimal way of achieving this collaboration is through the expertise of project managers, product owners, and business analysts who specialize in gathering requirements, documenting specifications, uncovering pain points, defining business key performance indicators [M1-4, M16, M18, C1-3, C10-11, C14, C16, C17]. Given the need for a wide variety of expertise, close collaboration of a multidisciplinary team is central to the successful design and development of Artificial Intelligence models for healthcare. Our proposed inception phase incorporates integration considerations into the *context understanding* phase of current Machine Learning model development lifecycles.

The Human considerations during this phase include identification of all stakeholders (clinicians, data scientists, and modeling experts) and clear problem definition [M17, M18, C1 & C2]. As part of defining the problem, it is important to identify existing human challenges with the problem (e.g., who is affected by the problem, how and why), any previous attempts at addressing the problem (e.g., what has been tried, what has worked, what failed and why), and ensuring a Machine Learning solution is suitable to the problem. Some potential strategies to achieve these objectives at this phase include stakeholder engagement, informal focus groups, and immersion in the problem space. Early involvement of stakeholders provides them with a greater understanding of the limitations of Machine Learning as well as investment in model evaluation, knowledge translation, integration, and monitoring (4, 25). Similarly, Machine Learning model developers and data

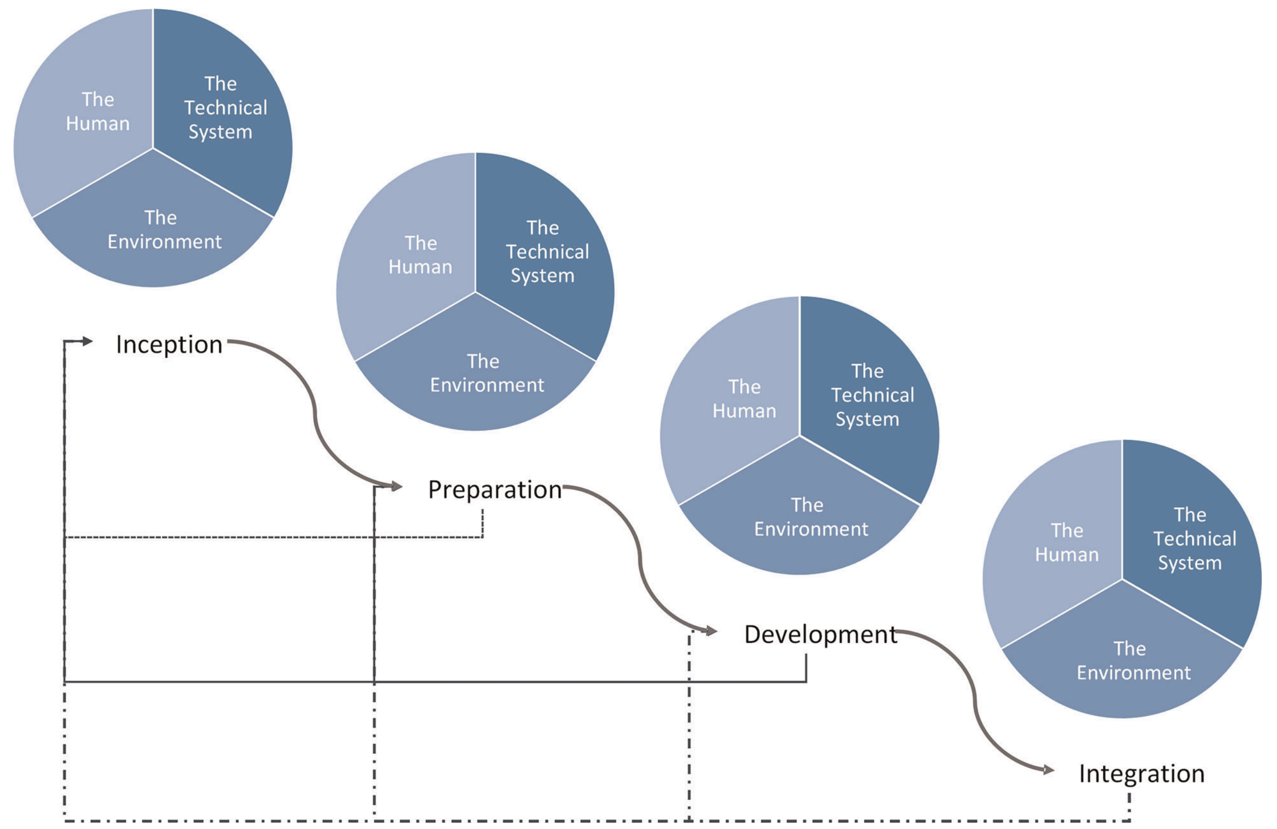


FIGURE 2
Proposed healthcare AI integration framework. Curved arrows show the progression through the integrated AI development process while the straight arrows show feedback from each phase into a previous phase. The framework begins in the top left with Inception and moves down and to the right, culminating in Integration.

scientists can also gain insight into the nature of the medical problem they are solving, as well as the unique features and properties of the data they would analyze and the environment they would be designing for.

Some *Environmental* considerations during this phase include an understanding of environmental constraints, the kinds of necessary hardware for the model training and operation, and data storage. This assessment of the environment can allow for more accurate estimates of cost as it relates to the clinical integration of the model which can facilitate estimations on feasibility of integration as well as future evaluation of balancing measures and project costs [C16]. Supplementary Table S1 illustrates the practical application of this phase of our proposed framework as it relates to the arrhythmia detection model.

Preparation

The *Technical System* related considerations during the preparation phase ensure relevant data are consistently, accurately, and reliably acquired and labelled for model training and evaluation [C4-7]. This is particularly important in clinical data as high-quality labels require the expertise of clinicians who are often restricted by availability. In addition, disagreements between multiple clinicians labeling the data can also introduce noise to the data and impact model training [C4-5, C7]. Preliminary analysis is also done on the "real" data to ensure it is suitable for the model as imagined [M3, M5, C1, C7] as well as ensuring any systematic bias or inaccuracies are identified and addressed [C8, C15]. This phase combines the current model development lifecycle phases of *data curation, data modeling* while also incorporating integration considerations.

Human considerations include the completion of a formal needs assessment, as well as cognitive and workflow analysis to identify specific needs and inefficiencies [M1, C2, C3]. User requirements are further defined and clarified in this phase through engagement with representatives from all stakeholders [M1, M18]. This can be achieved through cognitive task analysis, task analysis, workflow analysis, focus groups, interviews, and simulations. This establishes expectations which are calibrated during model development. Knowledge gaps among model users should be formally studied and identified to guide the development of a knowledge translation plan in the Development phase [C17]. This includes knowledge gaps in the clinical field for which a model is being designed for.

Environmental considerations in this phase include an evaluation of existing privacy and data security measures, as well as ethical and policy regulations that may need to be further developed to facilitate model integration and clinical utility. Supplementary Table S2 illustrates the practical application of this phase of our proposed framework as it relates to the arrhythmia detection model.

Development

During this phase, the *Technical System* (the model and its associated user interface) is developed through iterative design, testing, and statistical evaluations. This process should mimic that of agile software design (RAD model) with rapid modeling and testing of the model output evaluated against clinical gold standards and with real production datasets [M6, M7, M14, C7 & C11]. Other SDLC models can be more expensive, slower to iterate and develop, and present fewer opportunities for user and stakeholder engagement throughout the development cycles. This rapid and iterative development and testing would allow early and real-time feedback on the performance of the model [M8-10, M13, & M17] and generate performance matrices that can be used in testing with users to determine acceptable ranges of performance based on the clinical context of the application [M2, & M12]. The results of these assessments and the features of the model at each stage should be recorded to allow future auditing as well as the establishment of a track record and rigor that would be important in establishing clinician trust [M16]. Model output should also be reproducible through thoughtful selection of training and testing data instead of random sampling [M11]. These objectives can be realized through iterative retrospective studies as well as prospective silent trials that can be coupled with simulation testing to evaluate some of the *system-human* interaction considerations. If explainable Artificial Intelligence is required to achieve transparency of a model and facilitate its clinical integration, the explanations should also be developed during this time and evaluated together with the technical system itself for both accuracy and relevance, as well as usability and impact on decision making as described below [C12, C16]. Other means of achieving transparency include ensuring data properties (including pre-processing techniques), algorithmic properties, validation testing results, and other properties are well documented and disclosed to users as needed [C12]. These would be most relevant at the initial adoption phase of a model until its performance is experientially understood by users. The *production* phase of current model development lifecycles correlates to this new proposed phase.

The technical system is also refined based on considerations of the *Human* which involve a variety of formative and summative assessments that ensure user centered design and a robust knowledge translation strategy. Simulations and other human factors assessments should be used during this phase to evaluate the model's fit within the cognitive schema of users and their existing or proposed workflow. These investigations allow for further improvements to the model. Based on the stage of the model development, simulations of varying fidelity, from low fidelity tabletop activities to evaluate workflow and the model's user interface, to high fidelity in person simulations done in near-live environments with potentially real or realistic patient data and scenario can be conducted. Another opportune moment to understand relevant interactions is during silent trials where the model is run on real data, but its output is not made visible to the clinicians providing direct patient care. At the same time, the model can be made available to off-service clinician representatives from the previously defined user groups, and their interactions with the tool as well as its effect on their decision making, workflow, efficiency, accuracy, team dynamics, and much more can be studied and evaluated [C3]. The results of such a study can be fed back to the system developers to further optimize the system as well as used by policy makers and HFEs to refine the environment, training materials, education sessions, policies and much more. Depending on the complexity of the system and the fidelity of the simulation, these studies should be repeated until there is consistency and satisfaction in the performance of the SOS. It's important to note that through these simulations, broader user engagement can be achieved which would serve as a medium through which the stakeholders are made familiar with a tool that is under development as well as the ways in which it is being proposed for use clinically [M17-19, & C16].

The *Environmental* considerations during this phase ensure the development and security of software and hardware that facilitate optimal model operation within its intended environment. Infrastructure is also evaluated and optimized for various failure modes, inefficiencies, and instabilities. Necessary policies, regulations, and rules of engagement with the model, as well as ethical and privacy guidelines would also be developed and evaluated during this phase of model development [C13, C14]. Supplementary Table S3 illustrates the practical application of this phase of our proposed framework as it relates to the arrhythmia detection model.

Integration

The *Technical System* at this phase should be performing optimally for its intended problem, stakeholders, and healthcare context. During this phase, the model is launched for a prospective clinical evaluation [C10] and further refined based on live performance and user feedback [M19]. These prospective clinical evaluations should be subject to the peer review process to facilitate model adoption [C10]. Prospective studies should also be conducted to evaluate any long-term effects of model adoption (e.g., performance degradation secondary to practice changes) [C10] as well as any indications to suspect algorithmic biases [C14]. Finally, system monitoring should be ensued. This includes monitoring system metrics (e.g., server load, throughput, latency, etc.), input metrics (e.g., number of missing values, failed event detections, minimums, maximums, means, standard deviations, central frequencies, etc.), and output metrices (e.g., null predictions, model confusion, rate of change, etc.). It should also be evaluated for possible effects of data shift secondary to practice driven changes in data which may be due to the use of the model itself [M15, C9].

The *Human* considerations involve the launch of the developed knowledge translation plan prior to model deployment with ongoing just-in-time training when the model is clinically deployed. User engagement in the previous stages to develop knowledge translation strategies as well as evaluate model performance and usability would have allowed opportunities to calibrate user expectations to model capabilities and performance. Nevertheless, a formal knowledge translation process before the launch of the technical system and during the integration phase will further calibrate these expectations and establish a functional understanding of the technical system's operation and place within the existing workflow [M1, M17, C17]. Ongoing cost, workflow, and cognitive assessments are also leveraged as a feedback mechanism to further refine the overall system [C2].

The *Environmental* considerations during this phase, in addition to rolling out the developed and tested infrastructure as well as policies, procedures, privacy, and ethical considerations should also include an ongoing evaluation and refinement of these in practice. Supplementary Table S4 illustrates the practical application of this phase of our proposed framework as it relates to the arrhythmia detection model.

Upon the completion of this last phase, the Artificial Intelligence system is expected to be fully integrated into its intended healthcare environment and ready for clinical use.

Discussion

The healthcare environment is a complex SOS with multiple integrated systems from a wide variety of domains. Recognizing this complexity, our framework takes a systems engineering approach to Machine Learning model design for integration.

The interdisciplinary approach promoted by systems engineering ensures that the interactive components of a system are organized to achieve the purpose of the system (9, 10). For Machine Learning system design in the healthcare space, utilizing clinical, legal, ethical, and human factors expertise is as important as ensuring adequate Machine Learning, data science, and infrastructure expertise in the design and development process. Therefore, these aspects of a healthcare Machine Learning system are incorporated from inception and the involvement of experts from these different disciplines are expected throughout the development lifecycle of such systems in our framework. Involving project managers and business analysts, for example, can further strengthen the collaboration among different domain experts and facilitate the development of a well-designed product. End-user engagement is also emphasized in our framework as early and close engagement with the development of Machine Learning systems for healthcare will lead to improvements in system performance and fit. Greater transparency from this early and collaborative engagement, as well as user driven development of model explanations contribute to end-users' trust and adoption of Machine Learning.

As the number of integrated models for the healthcare domain increases and as users become increasingly familiar with Machine Learning models in healthcare, the need to integrate models designed in other clinical environments and institutions will increase. Our proposed framework could be used to evaluate these existing models for their fit into the new SOS. Issues with performance due to differences in data properties, fit in the workflow, usability, policies, environmental space, hardware requirements and others can be identified through the steps in this framework. Solutions to these issues can then be developed and tested until deemed adequate before the model is integrated into the new SOS.

Finally, the transformational opportunities that Machine Learning, and more broadly Artificial Intelligence, offer the field of medicine and healthcare, range from improving quality and efficiency in healthcare, improving accessibility, personalizing medicine, to advancing the field of medicine and healthcare. These models are fundamentally different to the current technology used in healthcare and they can not only move research in medicine from a hypothesis driven model to one that is data driven, but also modify clinical decision-making to be more data driven as well (47). The rigorous development and integration of these models is therefore crucial in maximizing their beneficial impact on healthcare.

Limitations

The framework that is proposed here aims at guiding the development of models for healthcare from their inception to their integration into the intended clinical space. For systems to be successful, they should also have a maintenance plan which would allow for ongoing modifications to optimize the system within its broader SOS in response to rare scenarios that may not have been encountered during the development phase as well as changes in the SOS that occur over time. For Machine Learning models, this is particularly relevant as the introduction of new technologies and changing practice patterns can have significant implications for the data based on which these models operate, potentially leading to a deterioration in model performance. The degradations in model performance should be continuously monitored and acted upon to ensure ongoing acceptable performance and reliability in their clinical application. To ensure due attention to integration without overlooking the important features in integration and maintenance, the detailed description of a maintenance phase for health care models was left out of this manuscript. We intend on exploring and developing a comprehensive maintenance phase in our future work. This will include guidelines on monitoring and maintaining longitudinal key performance indicators and model performance, as well as establishing thresholds for model retraining, workflow modifications, user re-training, and environmental modification (e.g., new policies, technology, and laws).

Conclusion

Artificial Intelligence models in healthcare are technical systems that need to be integrated into an existing system of systems that also includes the human and the environment. An integration engineering approach allows the creation of a pragmatic framework that we believe will both address the translation gap and inform and support regulatory approaches to Artificial Intelligence models in healthcare.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding author/s.

Author contributions

AA, PCL, DE, and MM closely collaborated to develop the presented framework. AG, SGoodfellow, WD, and RG in collaboration with DE developed the arrhythmia detection model and provided technical insight into the features of the framework pertaining to model development and evaluation. AG also provided additional overarching technical insight. AJ and DS in collaboration with AA, PCL, DE, and MM provided clinical insight into development of the framework while MC provided the ethical insight and considerations. SG provided insights pertaining to cognitive load. All authors contributed to the article and approved the submitted version.

Conflict of interest

MMcCradden is the John and Melinda Thompson Chair of Artificial Intelligence in Medicine and acknowledges funding from the SickKids Foundation relating to this role. All other authors have no conflicts of interest pertaining to this work to disclose.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2022.932411/full#supplementary-material.

References

1. IBM Cloud Education. Artificial intelligence (2020). Available at https://www.ibm.com/cloud/learn/what-is-artificial-intelligence (cited June 24, 2022).

2. FDA. FDA: What we do (2018). Available at https://www.fda.gov/about-fda/what-we-do (cited June 24, 2022).

3. Vincente K. *The human factor*. New York, USA: Routledge (2006).

4. Salvendy G. *Handbook of human factors and ergonomics*. 4th ed. New Jersey: John Wiley & Sons, Inc. (2012). 1–1736.

5. IEEE SA. About Us. Available at https://standards.ieee.org/about/ (cited June 24, 2022).

6. ISO. Standards. Available at https://www.iso.org/standards.html (cited June 24, 2022).

7. Rowe M. An Introduction to machine learning for clinicians. *Acad Med*. (2019) 94(10):1433–6. doi: 10.1097/ACM.0000000000002792

8. Akinsola JET, Ogunbanwo AS, Okesola OJ, Odun-Ayo IJ, Ayegbusi FD, Adebiyi AA. Comparative Analysis of Software Development Life Cycle Models (SDLC) (2020). 310–22. Available at http://link.springer.com/10.1007/978-3-030-51965-0_27

9. Rajabalinejad M, van Dongen L, Ramtahalsing M. Systems integration theory and fundamentals. *Saf Reliab*. (2020) 39(1):83–113. doi: 10.1080/09617353.2020.1712918

10. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *npj Digit Med*. (2018) 1(1):10–2. doi: 10.1038/s41746-018-0048-y

11. Kelly CJ, Karthikesalingam A, Suleyman M, Corrado G, King D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med*. (2019) 17(1):1–9. doi: 10.1186/s12916-019-1426-2

12. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med [Internet]*. (2019) 25 (1):30–6. doi: 10.1038/s41591-018-0307-0

13. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. (2019) 25(9):1337–40. doi: 10.1038/s41591-019-0548-6

14. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. *N Engl J Med*. (2019) 380(14):1347–58. doi: 10.1056/NEJMra1814259

15. Food US, Administration D. Software as a medical device (SAMD): clinical evaluation guidance for industry and food and drug administration staff. *FDA Guid*. (2017):1–32. Available at https://www.fda.gov/media/100714/download

16. Medical Device Coordination Group. Guidance on Qualification and Classification of Software in Regulation (EU) 2017/745—MDR and Regulation (EU) 2017/746—IVDR (2019):28. Available at https://ec.europa.eu/docsroom/documents/37581

17. ISO/IEC/IEEE. Systems and Software Engineering—System life cycle processes. ISO/IEC/IEEE 15288 (2015);1.

18. Friedenthal S, Moore A, Steiner R. Systems engineering overview. In: *A Pract Guid to SysML [Internet]*. Third. Elsevier (2015) p. 3–14. https://linkinghub.elsevier.com/retrieve/pii/B9780128002025000011

19. Buede DM. The engineering design of systems [Internet]. In: M Dennis, J Wiley, editors. *The engineering design of systems: models and methods*. Hoboken, NJ, USA: John Wiley & Sons, Inc. (2009). p. 1–48. http://doi.wiley.com/10.1002/9780470413791

20. Schlager KJ. Systenas engineering-key to modern development. *IEEE Trans Eng Manag*. (1956) EM-3(3):64–6. doi: 10.1109/IRET-EM.1956.5007383

21. Jamshidi M. System of systems engineering—new challenges for the 21st century. *IEEE Aerosp Electron Syst Mag*. (2008) 23(5):4–19. Available at http://ieeexplore.ieee.org/document/4523909/ doi: 10.1109/MAES.2008.4523909

22. Woo DM, Vicente KJ. Sociotechnical systems, risk management, and public health: Comparing the North Battleford and Walkerton outbreaks. *Reliab Eng Syst Saf*. (2003) 80(3):253–69. Available at https://linkinghub.elsevier.com/retrieve/pii/S0951832003000528 doi: 10.1016/S0951-8320(03)00052-8

23. Carayon P, Wooldridge A, Hoonakker P, Hundt AS, Kelly MM. SEIPS 3.0: human-centered design of the patient journey for patient safety. *Appl Ergon*. (2020) 84:103033. Available at https://linkinghub.elsevier.com/retrieve/pii/S000368701930239X doi: 10.1016/j.apergo.2019.103033

24. Rajabalinejad M. Incorporation of safety into design by safety cube. *Int J Ind Manuf Eng*. (2018) 12(3):476.

25. American Association for Medical Instrumentation. AAMI HE75: 2009(R 2018) | Human Factors Engineering—Design of Medical Devices (2018). Available at https://infostore.saiglobal.com/en-au/Standards/AAMI-HE75-2009-R-2018-517_SAIG_AAMI_AAMI_2628453/

26. Kneuper R. Sixty years of software development life cycle models. *IEEE Ann Hist Comput*. (2017) 39(3):41–54. doi: 10.1109/MAHC.2017.3481346

27. Shafiq S, Mashkoor A, Mayr-Dorn C, Egyed A. A literature review of using machine learning in software development life cycle stages. *IEEE Access*. (2021) 9:140896–920. doi: 10.1109/ACCESS.2021.3119746

28. Kumar M, Rashid E. An efficient software development life cycle model for developing software project. *Int J Educ Manag Eng*. (2018) 8(6):59–68. doi: 10.5815/ijeme.2018.06.06

29. Stephens R. *Beginning software engineering*. Indianapolis: John Wiley & Sons, Inc. (2015).

30. Kruse CS, Kristof C, Jones B, Mitchell E, Martinez A. Barriers to electronic health record adoption: a systematic literature review. *J Med Syst*. (2016) 40 (12):252–8. doi: 10.1007/s10916-016-0628-9

31. Mishra A. Amazon Machine learning. *Mach Learn AWS Cloud*. (2019):317–51. doi: 10.1002/9781119556749.ch15. Available at https://docs.aws.amazon.com/machine-learning/latest/dg/the-machine-learning-process.html

32. Microsoft Azure. What is the Team Data Science Process? (2022). Available at https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview (cited February 27, 2022).

33. Ferlitsch A. Making the machine: the machine learning lifecycle (2019). Available at https://cloud.google.com/blog/products/ai-machine-learning/making-the-machine-the-machine-learning-lifecycle (cited February 27, 2022).

34. Maskey M, Molthan A, Hain C, Ramachandran R, Gurung I, Freitag B, et al. Machine learning lifecycle for earth science application: a practical insight into production deployment. *Int Geosci Remote Sens Symp*. (2019):10043–6. doi: 10.1109/IGARSS.2019.8899031

35. Wan Z, Xia X, Lo D, Murphy GC. How does machine learning change software development practices? *IEEE Trans Softw Eng*. (2021) 47(9):1857–71. doi: 10.1109/TSE.2019.2937083

36. Shelley K, Cannesson M. Off-label use of clinical monitors: what happens when new physiologic understanding meets state-of-the-art technology. *Anesth Analg*. (2014) 119(6):1241–2. doi: 10.1213/ANE.0000000000000479

37. Antoniou T, Mamdani M. Evaluation of machine learning solutions in medicine. *CMJA*. (2021) 193(36):1425–9. doi: 10.1503/cmaj.210036

38. Karches KE. Against the iDoctor: why artificial intelligence should not replace physician judgment. *Theor Med Bioeth*. (2018) 39(2):91–110. doi: 10.1007/s11017-018-9442-3

39. He J, Baxter SL, Xu J, Xu J, Zhou X, Zhang K. The practical implementation of artificial intelligence technologies in medicine. *Nat Med*. (2019) 25(1):30–6. doi: 10.1038/s41591-018-0307-0

40. Hee Lee D, Yoon SN. Application of artificial intelligence-based technologies in the healthcare industry: Opportunities and challenges. *Int J Environ Res Public Health*. (2021) 18(1):1–18. doi: 10.3390/ijerph18010271

41. Aung YYM, Wong DCS, Ting DSW. The promise of artificial intelligence: a review of the opportunities and challenges of artificial intelligence in healthcare. *Br Med Bull*. (2021) 139(1):4–15. Available at https://academic.oup.com/bmb/article/139/1/4/6353269 doi: 10.1093/bmb/ldab016

42. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med*. (2021) 27(4):582–4. doi: 10.1038/s41591-021-01312-x

43. Wu E, Wu K, Daneshjou R, Ouyang D, Ho DE, Zou J. How medical AI devices are evaluated: limitations and recommendations from an analysis of FDA approvals. *Nat Med*. (2021) 27(4):582–4. doi: 10.1038/s41591-021-01312-x

44. Young AT, Amara D, Bhattacharya A, Wei ML. Patient and general public attitudes towards clinical artificial intelligence: a mixed methods systematic review. *Lancet Digit Heal*. (2021) 3(9):e599–611. doi: 10.1016/S2589-7500(21)00132-1

45. Hripcsak G, Albers DJ. Next-generation phenotyping of electronic health records. *J Am Med Informatics Assoc*. (2013) 20(1):117–21. doi: 10.1136/amiajnl-2012-001145

46. Demiris G, Oliver DP, Washington KT. Defining and analyzing the problem. *Behav Interv Res Hosp Palliat Care*. (2019):27–39. doi: 10.1016/B978-0-12-814449-7.00003-X

47. Zhu L, Zheng WJ. Informatics, data science, and artificial intelligence. *JAMA*. (2018) 320(11):1103. doi: 10.1001/jama.2018.8211

TYPE Original Research
PUBLISHED 16 August 2022
DOI 10.3389/fdgth.2022.929508

The silent trial - the bridge between bench-to-bedside clinical AI applications

Jethro C. C. Kwong[1,2†], Lauren Erdman[2,3,4†], Adree Khondker[5], Marta Skreta[3], Anna Goldenberg[2,3,4], Melissa D. McCradden[2,6,7,8], Armando J. Lorenzo[1,5] and Mandy Rickard[5]*

[1]Division of Urology, Department of Surgery, University of Toronto, Toronto, ON, Canada, [2]Temerty Centre for AI Research and Education in Medicine, University of Toronto, Toronto, ON, Canada, [3]Centre for Computational Medicine, The Hospital for Sick Children, Toronto, ON, Canada, [4]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, [5]Division of Urology, Department of Surgery, The Hospital for Sick Children, Toronto, ON, Canada, [6]Department of Bioethics, The Hospital for Sick Children, Toronto, ON, Canada, [7]Division of Clinical and Public Health, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada, [8]Genetics & Genome Biology, Peter Gilgan Centre for Research and Learning, Toronto, ON, Canada

Check for updates

OPEN ACCESS

EDITED BY
David Vidal,
Mayo Clinic, United States

REVIEWED BY
Aditya U. Kale,
University Hospitals Birmingham NHS Foundation Trust, United Kingdom
Samar Betmouni,
Sheffield Teaching Hospitals NHS Foundation Trust, United Kingdom
Jeffry Hogg,
Newcastle University, United Kingdom

*CORRESPONDENCE
Mandy Rickard,
mandy.rickard@sickkids.ca

†These authors share first authorship

SPECIALTY SECTION
This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health

RECEIVED 26 April 2022
ACCEPTED 01 August 2022
PUBLISHED 16 August 2022

CITATION
Kwong Jethro C. C., Erdman L, Khondker A, Skreta M, Goldenberg A, McCradden MD, Lorenzo AJ and Rickard M (2022) The silent trial - the bridge between bench-to-bedside clinical AI applications.
Front. Digit. Health 4:929508.
doi: 10.3389/fdgth.2022.929508

As more artificial intelligence (AI) applications are integrated into healthcare, there is an urgent need for standardization and quality-control measures to ensure a safe and successful transition of these novel tools into clinical practice. We describe the role of the silent trial, which evaluates an AI model on prospective patients in real-time, while the end-users (i.e., clinicians) are blinded to predictions such that they do not influence clinical decision-making. We present our experience in evaluating a previously developed AI model to predict obstructive hydronephrosis in infants using the silent trial. Although the initial model performed poorly on the silent trial dataset (AUC 0.90 to 0.50), the model was refined by exploring issues related to dataset drift, bias, feasibility, and stakeholder attitudes. Specifically, we found a shift in distribution of age, laterality of obstructed kidneys, and change in imaging format. After correction of these issues, model performance improved and remained robust across two independent silent trial datasets (AUC 0.85–0.91). Furthermore, a gap in patient knowledge on how the AI model would be used to augment their care was identified. These concerns helped inform the patient-centered design for the user-interface of the final AI model. Overall, the silent trial serves as an essential bridge between initial model development and clinical trials assessment to evaluate the safety, reliability, and feasibility of the AI model in a minimal risk environment. Future clinical AI applications should make efforts to incorporate this important step prior to embarking on a full-scale clinical trial.

KEYWORDS
dataset drift, bias, feasibility, stakeholder attitudes, artificial intelligence

Introduction

While artificial intelligence (AI) has gained much attention in healthcare, there is a pressing need for standardization and quality-control measures to ensure a safe and successful implementation into clinical practice. Premature deployment of machine learning (ML) models without rigorous external validation and governance can lead

TABLE 1 Major themes to explore during the silent trial before transitioning to the clinical trial phase. Each theme is associated with a suggested list of questions that should be considered.

Themes	Key questions
Dataset drift: Are there any changes between the training dataset and patients evaluated in the silent trial?[a]	1. Are there any changes as to how data are defined and collected? 2. Are there any changes to patient demographics, clinical settings, or unexpected events (i.e.: COVID-19) that would impact the patient population in which the model is applied? 3. Are there any changes in clinical practice such as indication, standard of care, or patient preference, that would influence the data being collected?
Bias: Was the model trained on a generalizable dataset to ensure fairness to all patients regardless of gender, race, etc.?	1. Which subset of patients benefit from the model? 2. Which subset of patients are harmed by the model?
Feasibility: Can the AI intervention be easily integrated within the existing clinical workflow?	1. How much time does it take for the end-user (i.e.: clinician) to input the necessary variables to generate a prediction? 2. How is the clinical workflow or duration of a clinic visit impacted with the use of the AI intervention? Importantly, does it slow down clinical workflow without a clear benefit? 3. Is the user interface simple enough to be used at point of care with minimal or no training? 4. Are the model predictions easy to understand? Are the model explanations easy to interpret? 5. How much computing resources or infrastructure are required to maintain the AI model at scale?
Stakeholder attitudes: Are there any concerns with respect to the use of AI to augment patient care?	1. Does the AI intervention facilitate patient counseling, decision-making, or treatment planning? 2. Are patients comfortable with the use of AI interventions to support their care? 3. What are the patient's priorities or goals of care regarding their condition and are they addressed by the AI intervention?

[a]Based on Finlayson et al. (13).

to discrepancies between reported and real-world performance, which may ultimately lead to patient harm. A recent example of this is the widely adopted Epic Sepsis Model that was found to have poor discrimination and calibration in predicting the onset of sepsis on external validation (1).

To mitigate these risks, several AI implementation pathways have been described (2, 3). We have previously outlined a 3-stage roadmap for the evaluation and validation of AI models into clinical care (4, 5), which has been implemented at scale at our institution. These phases include (1) exploratory model development, (2) a silent trial, and (3) prospective clinical evaluation. Several guidelines address the first and third phases to help standardize reporting, enhance reproducibility, and reliability of AI studies in healthcare (6–9). However, there has been limited discussion of the role of the silent trial, which evaluates the proposed model on patients in real-time, while the end-users (i.e., clinicians) are blinded to predictions such that they do not influence clinical decision-making. As shown in **Table 1**, this phase is essential to establishing feasibility and safety of AI models prior to proceeding with clinical evaluation where the model influences patient care.

The purpose of this article is to highlight the lessons learned from our experience in validating a previously developed model within the context of the silent trial. Here, we present the development of a classification model to predict obstruction in hydronephrotic kidneys of infants using ultrasound images. The current standard of care for infants with hydronephrosis, defined as swelling of one or both kidneys due to inadequate urinary drainage, involves serial ultrasounds typically every 3–6 months for several years. Patients may also undergo more invasive testing such as a diuretic renogram. While these investigations may provide useful information, the trade-off includes exposing patients to radioisotope and ionizing radiation as well as painful procedures such as venous canulation and urethral catheterization (10). Therefore, our aim was to develop an AI model that could reliably distinguish between self-resolving hydronephrosis vs. those that would ultimately require operative management based on initial kidney ultrasound images, thereby potentially reducing the number of invasive tests and expediting surgical interventions when necessary.

Using our model as a case study, we illustrate how issues related to dataset drift, bias, feasibility, and stakeholder attitudes were identified and addressed. This article is intended for clinicians and ML engineers wishing to gain a deeper understanding of the rationale behind the silent trial and provide insights as to how this phase serves as a bridge between initial model development and clinical trials assessment.

Materials and methods

Exploratory model development

We have previously developed a deep learning classification model to predict obstructive hydronephrosis in infants using still images from kidney ultrasound (11). Using sagittal and transverse images as inputs, the model would determine the probability of obstructive hydronephrosis and highlight areas of importance on the ultrasound images *via* GradCAM heatmaps (12). Obstructive hydronephrosis was defined by whether a patient ultimately required operative intervention to relieve the obstruction based on chart review. This tool was intended to be used at point-of-care to support clinical decision-making and patient counseling.

Silent trial

Following in silico/algorithmic validation, the AI model was prospectively validated in the silent trial from August to December 2020. During this period, the clinical team assessed and managed patients as per current standard of care. Concurrently, a separate research team recorded model predictions based on ultrasound images obtained at the time of the clinic visit. Additional patient demographics and the clinical decision to proceed with surgery were later collected. The clinical team was blinded to model predictions to avoid influencing clinical decision-making. This "Silent Trial 1" data was used to assess generalization of our initial model in a prospectively collected dataset (Figure 1). The results from Silent Trial 1 were then used to inform the refinement of the original model and data preprocessing steps. Once generalization was achieved on the Silent Trial 1 dataset, the original and Silent Trial 1 datasets were combined for model re-training. This updated model was then evaluated on another prospectively collected dataset, "Silent Trial 2". Model performance was characterized by area under the receiver-operating characteristic curve (AUROC) and area under the precision-recall curve (AUPRC), along with sensitivity and specificity found at a threshold set in the validation set targeting 90% sensitivity, based on consensus among the clinical expert group. The target of 90% sensitivity was chosen as false negatives would be particularly detrimental. Moreover, assessing model performance with a set threshold allowed us to test our model in a more real-world scenario of decision-making at a specific cut-off, rather than merely noting the separation of obstructed vs. non-obstructed cases.

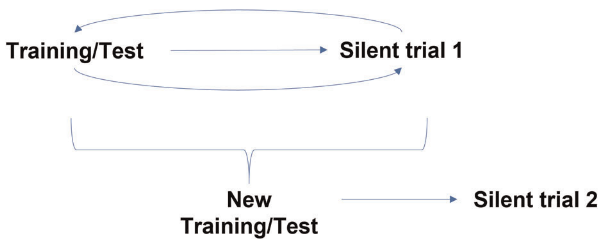


FIGURE 1
Silent trial workflow for model development. Initially, the model was trained and tested on a random 20% split of the initial dataset. Following successful generalization in this random split, the model was evaluated on new patients using prospectively collected data, Silent Trial 1. From this dataset, we identified any weaknesses in our model preventing it from generalizing successfully and adapted our initial model to overcome these limitations. Once the model generalized in this new set, the model was re-trained on both the initial and Silent Trial 1 datasets. This updated model was then tested on another prospectively collected data set, Silent Trial 2.

Given that AI in healthcare is still in its infancy, patient and family attitudes toward AI integration in their urologic care are not well understood. Therefore, it was essential to characterize patient perceptions about these tools to ensure that they were aligned with patient values and their role as a decision-support tool was clearly defined. To explore how patients and families would respond to the introduction of an AI tool into their care, we probed their initial thoughts and values through a standard post-visit follow-up questionnaire (Supplementary Table S1). This survey also sought to understand other patient priorities, such as the need for invasive testing, hospital visits, risks of infections, and renal impairment, however these were not the focus of this paper. Similarly, provider attitudes on the value of this AI intervention were assessed through clinical team meetings. We worked with multiple stakeholders in designing the user interface of our AI application. Feasibility was assessed by measuring the average time from starting the AI application to obtaining the probability of obstructive hydronephrosis based on user-uploaded ultrasound images.

Results

The initial training set contained 1,643 kidneys (1,456 non-obstructed/187 obstructed) from 294 patients (240 non-obstructed/54 obstructed) (Table 2). From a random test set of 20% drawn from the initial training set, the model achieved an AUROC of 90%, AUPRC of 58%, sensitivity of 92%, and specificity of 69% (Table 3, row 1). This model was then evaluated on the Silent Trial 1 dataset, which included 523 kidneys (387 non-obstructed/136 obstructed). This revealed a significant drop in performance with an AUROC of 50%, AUPRC of 26%, sensitivity of 100%, and specificity of 0% (Table 3, row 2). The following sections highlight how the silent trial enabled us to improve model performance and clinical utility by systematically examining the model with respect to dataset drift, bias, feasibility, and stakeholder attitudes.

Dataset drift

Through multidisciplinary discussions, we hypothesized several reasons for this change in performance including a shift in (1) age distribution, (2) distribution of laterality of obstructed kidneys, and (3) a change in processing of the input images (Figure 2). Indeed, patients included in the Silent Trial 1 dataset were younger (35 ± 39 vs. 61 ± 92 weeks, $p < 0.01$, Figure 2A), had predominantly right-sided obstructed kidneys (42 vs. 36%, $p < 0.01$, Figure 2B), and were visually different

TABLE 2 Baseline characteristics of each dataset.

Variable	Non-obstructed			Obstructed			Total		
	Training	Silent Trial 1	Silent Trial 2	Training	Silent Trial 1	Silent Trial 2	Training	Silent Trial 1	Silent Trial 2
Sex									
Male	981	326	530	138	104	69	1,119	430	599
Female	247	61	106	42	32	6	289	93	112
Age groups									
<2 years	1,025	359	561	171	128	71	1,196	487	632
2–5 years	143	28	72	9	6	0	152	34	72
>5 years	60	0	3	0	1	3	60	1	6
Ultrasound number									
1	403	127	214	69	46	28	472	173	242
2	316	110	184	50	39	24	366	149	208
3	248	74	130	34	24	11	282	98	141
4	161	39	63	19	12	8	180	51	71
5	112	16	32	8	5	2	120	21	34
6	84	11	10	3	3	1	87	14	11
7	63	6	2	4	3	1	67	9	3
8	37	3	1	0	2	0	37	5	1
9	18	1	0	0	1	0	18	2	0
10	13	0	0	0	1	0	13	1	0
11	1	0	0	0	0	0	1	0	0
Ultrasound Machine									
Philips	891	88	155	101	33	23	992	121	178
Samsung	34	59	125	2	21	17	36	78	125
Toshiba	448	229	347	69	48	32	517	277	379
GE	37	1	0	8	9	1	45	10	1
Acuson	23	0	0	2	0	0	25	0	0
ATL	17	0	0	5	0	0	22	0	0
Siemens	4	0	0	0	0	0	4	0	0
Outside	0	10	9	0	25	2	0	35	11
APD Group									
<6 mm	113	157	284	6	1	0	119	158	284
6–9 mm	119	92	150	6	10	5	125	102	155
9–14 mm	190	69	131	29	34	7	219	103	138
>14 mm	187	64	70	139	90	62	326	154	132
Not measured	847	5	1	7	1	1	854	6	2
Kidney view side									
Right	737	192	222	68	57	14	805	249	236
Left	719	195	414	119	79	61	838	274	475
Hydronephrosis side									
Right	673	126	83	56	52	13	729	178	96
Left	635	143	275	106	61	61	741	204	336
Bilateral	148	118	278	25	23	1	173	141	279
Overall observations	1,456	387	636	187	136	75	1,643	523	711
Overall unique patients	240	105	174	54	45	28	294	150	202

APD, anterior-posterior diameter.

TABLE 3 Iterative model performance.

Row	Train	Test	Model	AUROC	AUPRC	Sensitivity	Specificity
1	Original set	Random 20% from original set	Image only	0.90 (0.85, 0.95)	0.58 (0.39, 0.74)	0.92 (0.81, 1.0)	0.69 (0.63, 0.74)
2	Original set	Silent trial 1	Image only	0.50 (0.50, 0.50)	0.26 (0.21, 0.32)	1.00 (1.00, 1.00)	0.0 (0.0, 0.0)
3	Original set	Silent trial 1	Age and side covariates	0.51 (0.506, 0.52)	0.26 (0.22, 0.32)	1.00 (1.00, 1.00)	0.0 (0.0, 0.0)
4	Original set	Silent trial 1	Age-ablated	0.57 (0.55, 0.59)	0.28 (0.24, 0.35)	1.00 (1.00, 1.00)	0.005 (0.0, 0.01)
5	Original set	Silent trial 1	Side-ablated	0.54 (0.52, 0.55)	0.27 (0.22, 0.34)	1.00 (1.00, 1.00)	0.005 (0.0, 0.01)
6	Original set	Silent trial 1	Revised data prep, with covariates	0.85 (0.81, 0.88)	0.67 (0.58, 0.75)	0.98 (0.95, 1.00)	0.32 (0.27, 0.36)
7	Original set	Silent trial 1	Revised data prep, image only	0.84 (0.80, 0.88)	0.65 (0.57, 0.74)	0.99 (0.96, 1.00)	0.26 (0.22, 0.31)
8	Original set + silent trial 1	Silent trial 2	Revised data prep, with covariates	0.91 (0.88, 0.94)	0.52 (0.41, 0.64)	0.97 (0.93, 1.00)	0.54 (0.50, 0.57)
9	Original set + silent trial 1	Silent trial 2	Revised data prep, image only	0.92 (0.88, 0.95)	0.52 (0.41, 0.64)	0.99 (0.95, 1.00)	0.52 (0.48, 0.56)

Values reflect performance of data in the Test column. Model formulation described in the Model column, indicating iterative experiments performed to rescue Silent trial performance. Sensitivity and specificity thresholds set in validation set targeting 90% sensitivity.

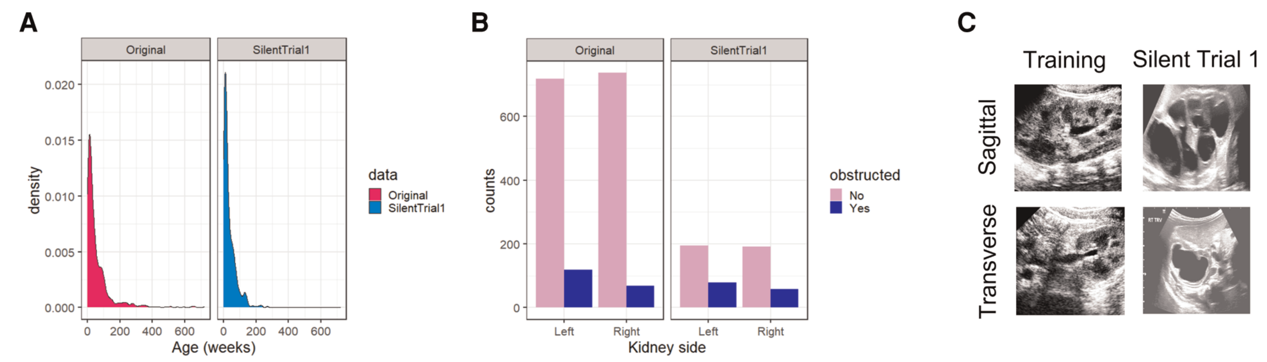


FIGURE 2
Dataset drift between our original training set and Silent Trial 1. (A) The shift in age to younger individuals in the Silent Trial 1 dataset. (B) The shift between left and right-sided kidneys in which a larger proportion of right-sided obstructed kidneys were found relative to the left in the Silent Trial 1 set. (C) The qualitative shift in images despite the same cropping and normalization procedures for both datasets.

even following the same preprocessing steps (Figure 2C). Therefore, we postulated that these differences may explain the precipitous drop in model performance.

To overcome the limitations of the original model in the Silent Trial 1 dataset, we adapted the original model to incorporate kidney laterality and patient age as covariates to adjust for the dataset drift (Figure 3). With this approach, we found a minor improvement in AUROC, although other performance metrics remained unchanged (Table 3, row 3). We then ablated each covariate by setting either all age or laterality values to zero to evaluate the degree to which each covariate impacted the model's performance, with the hypothesis that one may be more impactful than the other. This procedure resulted in a small but significant increase in model performance for each ablation (Table 3, rows 4–5).

We next turned to image preprocessing and found that the original dataset included processed jpeg files, whereas the Silent Trial 1 dataset included either unprocessed or processed png files. We first experimented with merely passing these images through the same preprocessing steps and reading them into the model. However, this had clearly not addressed the shift in image formatting. Therefore, we experimented with adding the additional step of saving our newly processed data as jpegs files and re-reading them into the model in the same format. This approach led to a tremendous boost in performance on the Silent Trial 1 dataset, with an AUROC of 85% for the model with covariates and 84% for the image-only model (Table 3, rows 6–7).

After addressing these dataset drift issues, we evaluated these updated models on a third dataset, Silent Trial 2, to confirm the generalizability of this approach and assess if covariates should continue to be included. We found that these models do indeed perform well on the Silent Trial 2 dataset, with an AUROC of 91% for the model with covariates and 92% for the image-only model (Table 3, rows 8–9).

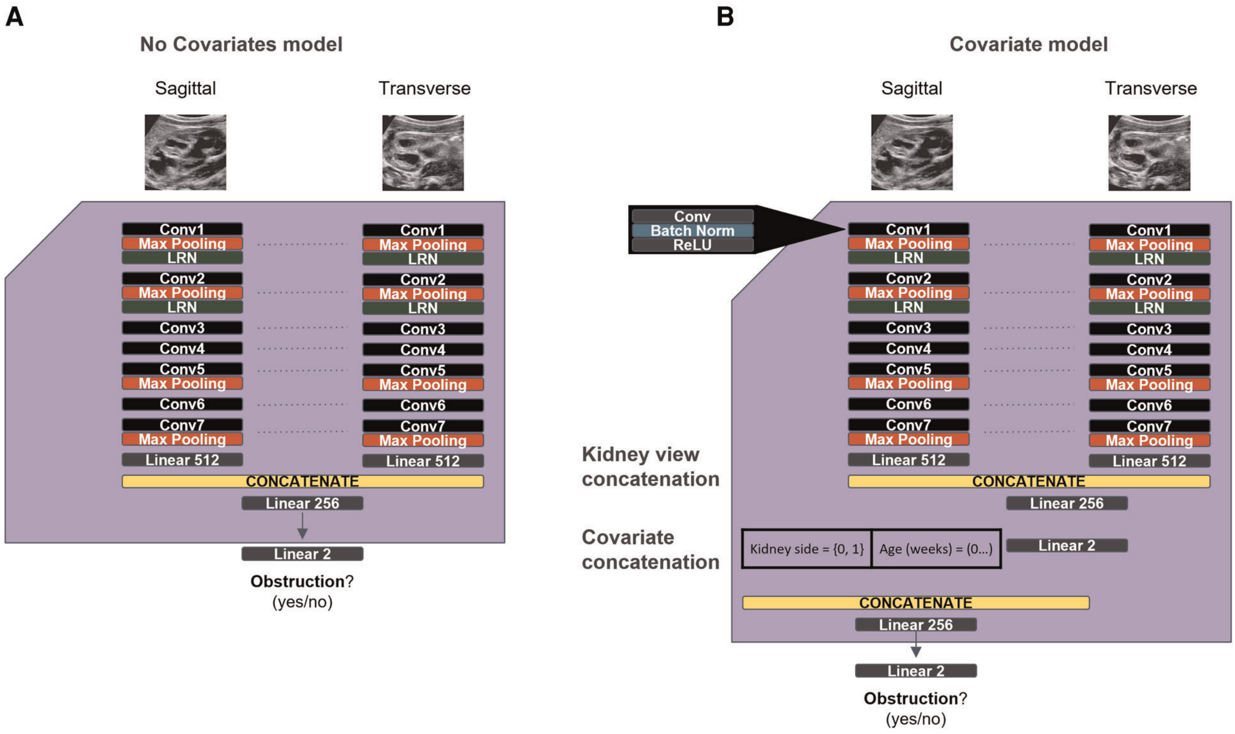


FIGURE 3
Original and updated models used to overcome dataset drift. (A) The original model used from the initial dataset. (B) Updated model with covariates for age and kidney laterality, with the goal of overcoming the generalization failure observed on the Silent Trial 1 dataset.

Bias

Next, we conducted a bias assessment of our model to ensure there were no substantial differences in performance when stratified by clinically relevant subgroups including sex, side of hydronephrosis, ultrasound machine, and patient postal code (Table 4). We find in all cases >90% sensitivity for each subgroup, therefore supporting the overall safety of our model. Specificity is far more variable, however in all cases, we find it >50%, therefore every group would benefit from safe and effective streamlined care with this model.

Feasibility

To ensure that the AI intervention was appropriate for routine clinical use, we considered whether the application was simple-to-use and minimally disruptive to the existing workflow. Feasibility was assessed by diverse stakeholders including clinicians, nurse practitioners, trainees, computer scientists, web developers, and patient representatives. The user interface for the AI application was developed using an iterative process involving all stakeholders to simplify instructions, improve clinical utility, and protect patient confidentiality (Figure 4). The average time to generate a model prediction from start-to-finish without prior training was less than one minute. Model output is saved locally within the computer that the program runs on and is analyzed without sending any data over the internet, therefore data and patient-specific findings remain confidential and secure.

Stakeholder attitudes

Understanding the views and perspectives of patients and providers were essential to ethical integration of the AI intervention. From the provider's perspective, the clinical team felt that this intervention would potentially augment their clinical care by identifying patients at risk of requiring surgical intervention for their hydronephrosis. These opinions were aligned with the potential benefits previously outlined by the clinical team during the model development phase (11). It would also provide useful clinical decision support without adding significant time to each patient visit.

A questionnaire on the use of AI in clinical care was distributed to patients and their families after clinic visits to explore whether they would be open to consenting to use of an AI intervention and if they felt it could address their primary concerns. Out of 44 respondents, 34 (77%) prioritized knowing

TABLE 4 Bias assessment of our final AI model.

Variable	AUROC	AURPC	Sensitivity	Specificity
Sex				
Male	0.91 (0.87, 0.94)	0.52 (0.42, 0.65)	0.97 (0.93, 1.00)	0.53 (0.49, 0.57)
Female	0.96 (0.91, 1.00)	0.38 (0.12, 0.80)	1.00 (1.00, 1.00)	0.59 (0.50, 0.68)
Side of hydronephrosis				
Left	0.88 (0.84, 0.93)	0.57 (0.45, 0.72)	0.97 (0.92, 1.00)	0.48 (0.43, 0.53)
Right	0.96 (0.91, 0.99)	0.61 (0.39, 0.86)	1.00 (1.00, 1.00)	0.60 (0.50, 0.71)
Both	0.98 (0.96, 0.99)	0.08 (0.05, 0.30)	1.00 (1.00, 1.00)	0.58 (0.52, 0.63)
Ultrasound machine				
Philips	0.89 (0.83, 0.95)	0.50 (0.31, 0.71)	0.96 (0.84, 1.00)	0.53 (0.46, 0.62)
Samsung	0.92 (0.86, 0.96)	0.50 (0.30, 0.71)	1.00 (1.00, 1.00)	0.58 (0.50, 0.66)
Toshiba	0.93 (0.86, 0.97)	0.53 (0.39, 0.72)	0.97 (0.90, 1.00)	0.53 (0.48, 0.58)
Postal code				
K	1.00 (1.00, 1.00)	0.86 (0.67, 0.91)	1.00 (1.00, 1.00)	0.50 (0.24, 0.82)
L	0.90 (0.85, 0.95)	0.49 (0.37, 0.65)	0.95 (0.89, 1.00)	0.58 (0.52, 0.63)
M	0.91 (0.86, 0.97)	0.57 (0.35, 0.76)	1.00 (1.00, 1.00)	0.50 (0.45, 0.56)
N	NA	NA	NA	0.75 (0.25, 1.00)
P	1.00 (1.00, 1.00)	0.86 (0.00, 0.91)	1.00 (1.00, 1.00)	0.57 (0.24, 0.89)

Performance of our model was stratified by sex, side of hydronephrosis, ultrasound machine, and postal code in our Silent Trial 2 set.

whether their child would require surgery as the most important, which was aligned with the primary objective of the AI intervention. Majority of respondents (68%) supported the use of AI in their care, while those who did not cited concerns that it would replace the physician-patient interaction or insufficient knowledge regarding AI itself. As a result, this questionnaire helped identify areas to educate patients and their families regarding how the AI intervention would act as a clinical adjunct to facilitate personalized and data-driven care.

Discussion

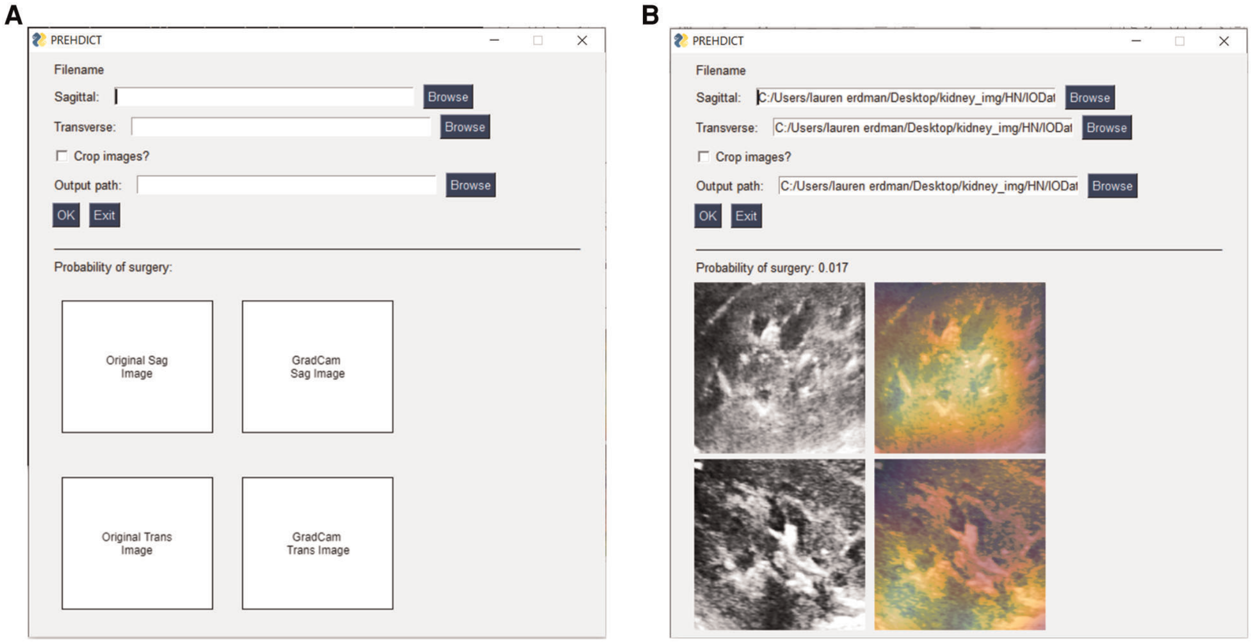
AI integration in healthcare is growing exponentially with a diverse range of applications, from aiding diagnosis and prognosis, to supporting treatment planning and patient counseling. As more AI applications move into the clinical space, researchers have an ethical obligation to evaluate these interventions in a minimal risk environment to ensure their safety and efficacy. This is the primary motivation behind the silent trial. Here, we demonstrate the iterative changes applied to our predictive model for obstructive hydronephrosis and the resulting improvements in its accuracy and generalizability.

Why is a silent trial warranted?

AI models trained on retrospective data alone cannot reliably function in real-world clinical settings as they are prone to dataset drift, which may include variations in how data is defined and collected, or potential changes in the standard of care if training cohorts span long periods of time (13). Use of real-time data may present additional challenges such as delays in preprocessing data or incomplete data at a given time-point. Other considerations include establishing a decision pathway and legal framework (14). A silent trial also facilitates an assessment of bias to ensure social disparities are not accentuated by the model. In this study, failure to adequately assess and account for bias may result in overtreatment of certain patient subgroups due to an inappropriately high predicted risk of obstructive hydronephrosis. We compared model performance with respect to sex, side of hydronephrosis, ultrasound machine, and patient postal code. Taken together, the silent trial enables clinicians and researchers to explore these issues in-depth without putting patients at risk of unvalidated predictions (15).

How did the silent trial improve the applicability of our model?

While the first iteration of our model demonstrated excellent discriminative capability on a retrospective exploratory dataset (AUROC 0.90), it performed poorly on real-time data (AUROC 0.50). By validating our model in a silent trial instead of a clinical trial setting, we were able to recognize this performance drop without subjecting patients to unnecessary harm due to misclassification. Through careful



A

Sagittal:

Transverse:

□ Crop images?

Output path:

Probability of surgery:

B

Sagittal: C:/Users/lauren erdman/Desktop/kidney_img/HN/IODat

Transverse: C:/Users/lauren erdman/Desktop/kidney_img/HN/IODat

□ Crop images?

Output path: C:/Users/lauren erdman/Desktop/kidney_img/HN/IODat

Probability of surgery: 0.017

FIGURE 4
User-interface for image-only model. A basic user-interface was developed to allow clinicians and researchers who are not computer scientists to test the model. (A) The user-interface with no data input, in which a user can specify a sagittal and transverse ultrasound image file of the kidney, along with an option for the program to further crop the image and where to save the output. The lower-half of the interface is blank at this point, as it will display the uploaded images. (B) This view now shows the user-interface once the model has run. This displays the probability of surgery, the original input images following the preprocessing procedure, and a gradient-based class activation maps to the image to indicate which part of the image is most important for the prediction.

consideration of dataset drift, potential sources of bias, and inclusion of other clinically relevant features, the model accuracy improved on the Silent Trial 1 dataset (AUROC 0.85) and remained robust when applied to the Silent Trial 2 dataset (AUROC 0.91). Another benefit of the silent trial is the potential to reveal discrepancies between the AI model and the current standard of care, which may highlight opportunities for quality improvement and promote hypothesis generation.

The silent trial also enables investigators to evaluate whether the proposed AI model is appropriate for real-world clinical applications. In contrast to performance evaluations which look at objective metrics, a feasibility assessment helps ensure adequate buy-in from all stakeholders. This is an essential consideration because even the most accurate AI model cannot provide meaningful clinical benefit if it is too time-consuming, difficult to use, not clinically relevant, or not endorsed by patients and physicians. In the present study, we identified nearly one-third of patients and families who were hesitant regarding the use of AI interventions to support patient care. Chew et al. found that patient concerns about AI integration were primarily attributed to a lack of trust in data privacy, patient safety, maturity of AI interventions, and risk of complete automation of their care (16). These findings underscore the need for more patient engagement prior to recruiting patients for an AI clinical trial and identifies key issues for patient education (e.g., AI will not replace their clinician, patients will still be seen by clinicians, etc.). Therefore, incorporating both patient and provider feedback into the design process can help build trust and strengthen the partnership between developers and end-users (17). Similarly, measuring patient-reported outcomes and health systems benefits in conjunction with traditional performance metrics may provide a more holistic assessment of how AI models may improve clinical practice (18). For example, a ML-based model to screen urine samples was accepted by providers because of the significant time and cost-savings without compromising care (19). Overall, the silent trial can not only refine model performance, but also facilitate a transition into clinical practice and better tailor a prospective clinical evaluation (3, 20).

Tips to successfully implement the silent trial

Several factors outlined below are vital to the success of implementing a silent trial. The clinical team should manage sufficient patient volumes for the proposed clinical question to accrue enough patient data for the silent trial within a reasonable timeframe. The research team should have appropriate AI expertise to adequately assess the proposed model across the four themes of the silent trial. Strong partnerships between the clinical and research teams are essential and infrastructure should be established to facilitate collaboration and regular meetings between the two groups. The host institution should also be capable of storing data and the final trained model. Finally, the project should be championed by a clinical expert who can secure funding and advocate for the implementation of the AI tool into clinical practice.

Limitations

Several limitations of this study merit discussion. The outcome of interest (obstructive hydronephrosis) was defined based on whether patients underwent surgical intervention. However, this may vary based on patient preferences, surgeon clinical judgement, and changes in clinical practice guidelines over time. Serial ultrasounds for each patient and kidney were treated as independent samples, therefore additional prognostic information from changes across serial ultrasounds may be lost. However, we felt that a rapid, point-of-care tool using single ultrasound images would be more beneficial in most clinical settings. Finally, our assessment of patient perspectives on implementation of our AI tool was based on an unvalidated questionnaire due to resource and time constraints. Future work can explore the use of validated surveys and the impact of our AI tool on patient reported outcomes.

Conclusion

Here, we highlight our experience with the silent trial, using an AI-based classification model for hydronephrosis as an illustrative example. This phase enables stakeholders to audit and report on issues related to dataset drift, bias, feasibility, and stakeholder attitudes. These are important considerations which must be made to ensure the safety, reliability, and feasibility of AI models in real-world clinical practice. Future clinical applications of AI should make efforts to demonstrate and reflect on model changes using this process.

Data availability statement

The data is not publicly available, since all research or research-related activities that involve an external party may require, at the discretion of The Hospital for Sick Children, Toronto, Canada, a written research agreement in order to define the obligations and manage the risks.

Ethics statement

The studies involving human participants were reviewed and approved by The Hospital for Sick Children. Written informed consent from the participants' legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

JCCK, Conceptualization, Writing - Original Draft/Review & Editing, Visualization; LE, Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Review & Editing, Visualization; AK, Writing - Review & Editing; MS, Writing - Review & Editing; AG, Writing - Review & Editing, Supervision; MDM, Writing - Review & Editing; AJL, Writing - Review & Editing, Supervision; MR, Writing - Review & Editing, Supervision. All authors contributed to the article and approved the submitted version.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2022.929508/full#supplementary-material.

References

1. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med*. (2021) 181(8):1065–70. doi: 10.1001/jamainternmed.2021.2626

2. Sendak MP, D'Arcy J, Kashyap S, Gao M, Nichols M, Corey KM, et al. A Path for Translation of Machine Learning Products into Healthcare Delivery. In (2020).

3. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. (2019) 25(9):1337–40. doi: 10.1038/s41591-019-0548-6

4. McCradden MD, Stephenson EA, Anderson JA. Clinical research underlies ethical integration of healthcare artificial intelligence. *Nat Med*. (2020) 26:1325–6. doi: 10.1038/s41591-020-1035-9

5. McCradden MD, Anderson JA, Stephenson E A, Drysdale E, Erdman L, Goldenberg A, et al. A research ethics framework for the clinical translation of healthcare machine learning. *Am J Bioeth*. (2022) 22(5):8–22. doi: 10.1080/15265161.2021.2013977

6. Kwong JCC, McLoughlin LC, Haider M, Goldenberg MG, Erdman L, Rickard M, et al. Standardized reporting of machine learning applications in urology: the STREAM-URO framework. *Eur Urol Focus*. (2021) 7(4):672–82. doi: 10.1016/j.euf.2021.07.004

7. Vasey B, Clifton DA, Collins GS, Denniston AK, Faes L, Geerts BF, et al. DECIDE-AI: new reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med*. (2021):1–2.

8. Liu X, Cruz Rivera S, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. (2020) 26:1364–74. doi: 10.1038/s41591-020-1034-x

9. Cruz Rivera S, Liu X, Chan AW, Denniston AK, Calvert MJ; SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Lancet Digit Health*. (2020) 2(10):e549–60. doi: 10.1016/S2589-7500(20)30219-3

10. Jacobson DL, Flink CC, Johnson EK, Maizels M, Yerkes EB, Lindgren BW, et al. The correlation between serial ultrasound and diuretic renography in children with severe unilateral hydronephrosis. *J Urol*. (2018) 200(2):440–7. doi: 10.1016/j.juro.2018.03.126

11. Erdman L, Skreta M, Rickard M, McLean C, Mezlini A, Keefe DT, et al. *Predicting obstructive hydronephrosis based on ultrasound alone*. In: *Lecture notes in computer science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer Science and Business Media Deutschland GmbH (2020). p. 493–503.

12. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: visual explanations from deep networks via gradient-based localization. *Int J Comput Vis*. (2020) 128(2):336–59. doi: 10.1007/s11263-019-01228-7

13. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med*. (2021) 385:283–6. doi: 10.1056/NEJMc2104626

14. Binkley CE, Green BP. Does intraoperative artificial intelligence decision support pose ethical issues? *JAMA Surg*. (2021). doi: 10.1001/jamasurg.2021.2055. [Epub ahead of print]

15. McCradden MD, Anderson JA, Zlotnik Shaul R. Accountability in the machine learning pipeline: the critical role of research ethics oversight. *Am J Bioeth*. (2020) 20(11):40–2. doi: 10.1080/15265161.2020.1820111

16. Chew HSJ, Achananuparp P. Perceptions and needs of artificial intelligence in health care to increase adoption: scoping review. *J Med Internet Res*. (2022) 24 (1):e32939. doi: 10.2196/32939

17. Yang Q, Steinfeld A, Zimmerman J. *Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes*. In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. (2019). p. 1–11.

18. Tasian GE, Ellison JS. The surgical improvement cycle: improving surgical outcomes through partnerships and rigor. *J Urol*. (2021) 205:1554–6. doi: 10.1097/JU.0000000000001626

19. Burton RJ, Albur M, Eberl M, Cuff SM. Using artificial intelligence to reduce diagnostic workload without compromising detection of urinary tract infections. *BMC Med Inform Decis Mak*. (2019) 19(1):171. doi: 10.1186/s12911-019-0878-9

20. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. *Br Med J*. (2015) 350:h391. doi: 10.1136/bmj.h391

TYPE Review
PUBLISHED 19 August 2022
DOI 10.3389/fdgth.2022.939292

Clinical deployment environments: Five pillars of translational machine learning for health

Steve Harris[1,2*†], Tim Bonnici[1,2†], Thomas Keen[1], Watjana Lilaonitkul[1], Mark J. White[4] and Nel Swanepoel[3]

[1]Institute of Health Informatics, University College London, London, United Kingdom, [2]Department of Critical Care, University College London Hospital, London, United Kingdom, [3]Centre for Advanced Research Computing, University College London, London, United Kingdom, [4]Digital Healthcare, University College London Hospital, London, United Kingdom

OPEN ACCESS

EDITED BY
Karandeep Singh,
University of Michigan, United States

REVIEWED BY
Manisha Mantri,
Center for Development of Advanced Computing (C-DAC), India
Patrick Lyons,
Washington University in St. Louis, United States
Jean Feng,
University of California, San Francisco, United States

*CORRESPONDENCE
Steve Harris
steve.harris@ucl.ac.uk

†These authors share first authorship

SPECIALTY SECTION
This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health

RECEIVED 09 May 2022
ACCEPTED 25 July 2022
PUBLISHED 19 August 2022

CITATION
Harris S, Bonnici T, Keen T, Lilaonitkul W, White MJ and Swanepoel N (2022) Clinical deployment environments: Five pillars of translational machine learning for health. Front. Digit. Health 4:939292.
doi: 10.3389/fdgth.2022.939292

Machine Learning for Health (ML4H) has demonstrated efficacy in computer imaging and other self-contained digital workflows, but has failed to substantially impact routine clinical care. This is no longer because of poor adoption of Electronic Health Records Systems (EHRS), but because ML4H needs an infrastructure for development, deployment and evaluation within the healthcare institution. In this paper, we propose a design pattern called a Clinical Deployment Environment (CDE). We sketch the five pillars of the CDE: (1) real world development supported by live data where ML4H teams can iteratively build and test at the bedside (2) an ML-Ops platform that brings the rigour and standards of continuous deployment to ML4H (3) design and supervision by those with expertise in AI safety (4) the methods of implementation science that enable the algorithmic insights to influence the behaviour of clinicians and patients and (5) continuous evaluation that uses randomisation to avoid bias but in an agile manner. The CDE is intended to answer the same requirements that bio-medicine articulated in establishing the translational medicine domain. It envisions a transition from "real-world" data to "real-world" development.

KEYWORDS
translational medicine, machine learning, health informatics, ML-Ops, safety, artificial intelligence

Introduction

Bold claims and huge investments suggest Machine Learning (ML) will transform healthcare (1). High impact publications showcase precision models that predict sepsis, shock, and acute kidney injury (2–4). Outside healthcare, tech titans such as AirBnB, Facebook, and Uber create value from ML despite owning "no property, no content and no cars" (5). Inspired by this, and very much aware of the flaws and unwarranted variation in human decision making (6), government and industry are now laying heavy bets on ML for Health (ML4H) (7, 8).

Widespread adoption of electronic health records (EHR) might be thought a sufficient prerequisite for this ambition. Yet while EHR adoption is growing at pace

(9), those ML4H models that have reached the market rarely use the EHR. They are instead embedded in isolated digital workflows (typically radiology) or medical devices (10). Here the context of deployment is static and self-contained (imaging), or fully specified (devices), and translation has proved easier to navigate.

In contrast, the EHR is in constant flux. Both the data and the data model are updating. New wards open, staffing patterns are adjusted and from time to time major incidents (even global pandemics) disrupt everything. There are multiple interacting users, and eventually there will be multiple interacting algorithms, and organizations will face the ML equivalent of poly-pharmacy (11). Algorithms will require stewards (12). Whilst the aforementioned high impact prediction models are developed on real-world data, this is not the same as real-world development. Data are either anonymized and analyzed offline, or moved out of the healthcare environment into an isolated Data Safe Haven (DSH) [also known as Trusted Research Environment (TRE)] (13). This separation is the first fracture leading to the oft-cited AI chasm (14) leaving the algorithms stranded on the laboratory bench.

A future that sees ML4H generate value from the EHR requires an alternative design pattern. TREs excel at meeting the needs of population health scientists but they do not have the full complement of features required to take an ML4H algorithm from bench-to-bedside. Using drug development as an an analogy, a TRE is custom made for drug discovery not translational medicine (15).

In this paper, we describe the functional requirements for a Clinical Deployment Environment (CDE) for translational ML4H. These requirements map closely to the classical components of translational medicine, but differ in that algorithms will require ongoing stewardship even after a successful deployment. The CDE is an infrastructure that manages algorithms with the same regard that is given to medicines (pharmacy) and machines (medical physics). Moreover, the value of ML4H will not just be from externally developed blockbuster models, but will also derive from specific and local solutions. Our vision of a CDE therefore enables both *development* and *deployment*.

Our CDE is supported by five pillars:

1. Real World Development
2. ML-Ops for Health
3. Responsible AI in practice
4. Implementation science
5. Continuous evaluation

We describe these pillars below alongside figures and vignettes reporting early local experience in our journey building this infrastructure.

1. Real world development

Real-world data (RW-Data) means the use of observational data at scale augmented by linking across multiple data sources to generate insights simply not available from isolated controlled clinical trials (16). The FDA uses data from tens of millions of patients in its Sentinel programme to monitor drug safety, and the OpenSafely programme in the UK generated impactful insights into COVID-19 within the first few months of the global pandemic (17).

Given the sensitive nature of health data, these initiatives depend on expanding investment into TREs (18). TREs are an example of “data-to-modeler” (DTM) designs where data flows from source (primary, secondary, social care and elsewhere) to a separate, secure landing zone. Here research teams write the code to link, clean and analyze the data. Derived insights eventually return to the bedside through clinical guidelines and policy. To date, DTM is also the dominant design pattern in ML4H but this approach is fundamentally flawed.

It is flawed because it imposes a separation between the modeller and the end-user. ML4H is not concerned with better guidelines or policy but with better operational and clinical decision making. This requires the practitioner to work alongside the end-user because excellent offline model performance provides no guarantee of bedside efficacy. Algorithms with inferior technical performance may even provide greater bedside utility (19, 20). An inverted “modeler-to-data” (MTD) paradigm was initially proposed to reduce privacy concerns (data are no longer copied and shared but analyzed *in situ* (21)), but we see important additional value in that it forces “real-world development” (RW-Dev) and enables the end-user to work with the modeler in rapid-cycle build-test-learn loops. This first pillar of the CDE is the equivalent of an *internal* TRE *within* the healthcare institution (21).

RW-Dev has four functional sub-requirements that distinguish it from a TRE. (1) Firstly, data updates must match the cadence of clinical decision making. For most inpatient and acute care pathways, decisions are in real-time (minutes or hours) at the bedside or in the clinic. (2) Secondly, development using live data must be sandboxed and so the clinical system responsible for care delivery is protected (3) thirdly, privacy must be managed such that teams are able to develop end-user applications that inevitably display patient identifiable information (PII) alongside the model outputs: an anonymous prediction is of little use to a clinician. (4) Fourthly, attention must be paid to developer ergonomics. Where development and deployment steps are separated physically (the TRE paradigm) or functionally (*via* different languages and technologies), ownership is often split between two different teams. One team prepares the raw data and develops the model, and another prepares the live data and deploys the model. We argue instead that the same team should be able develop *and* deploy. This should accelerate iteration, reduce cost and increase quality (22).

We illustrate this idea with a description of our local real-world development platform in **Figure 1**, and provide an extended description in the **Supplementary Material**.

2. ML-OPS (for health)

Hitherto in ML4H, the data and the algorithm have been the "celebrity couple". State-of-the-art models trained on RW-Data deliver high profile publications (3, 4). But only a tiny handful (fewer than 10 studies in a recent high quality systematic review of nearly 2000 ML4H publications (23)), were prospectively implemented. The standard offline "data-to-modeler" (DTM) paradigm described above incurs a significant but "hidden technical debt" that includes configuration, data collection and verification, feature extraction, analysis and process tools, compute and storage resource management, serving infrastructure, and monitoring (24). In fact, the code for the underlying ML model is estimated to be at most 5% of the total code with the other 95% as additional code to make the system work. "Glue-code", "pipeline jungles", and "dead experimental codepaths" are some of the anti-patterns that make the transition into production costly and hazardous.[1]

Agencies such as the FDA, EMA, and MHRA are working toward safety standards for AI and machine learning, but the majority of these efforts derive from medical devices regulation. Treating Software as a Medical Device (SaMD) is appropriate where the algorithms operate within a constant and predictable environment (e.g. code embedded within a cardiac pacemaker). But, as already argued, ML4H models working with the EHR are likely to find themselves operating in a significantly more complex landscape. This inconstant environment where algorithms themselves may only have temporary utility has parallels to the commercial environment exploited so successfully by the tech giants.

These companies have cultivated an approach to model deployment called "ML-Ops". This combines the practices of "DevOps" (a portmanteau of Software Development plus IT operations) (22) that focuses on the quality and speed with which software moves from concept to production, with robust data engineering and machine learning. A typical ML-Ops system monitors raw input data, checks for distribution drift, provides a feature store to avoid train/serve skew and facilitate collaboration between teams, and maintains an auditable and monitored model repository (26). We present a prototype implementation interacting with the EHRS in **Figure 2** (called FlowEHR).

This constant adjustment of algorithms based on their continuously measured quality and performance needs a workforce as well as a technology stack. Just as the safe delivery of medicines to the bedside is the central activity of a hospital pharmacy team, the safe delivery of algorithms will require the development of similarly skilled and specialized practitioners, and we should expect to see clinical ML-Ops departments in the hospital of the future. Others have made similar proposals and labeled this as "algorithmic stewardship" or "AI-QI" (12, 27). Similarly, the FDA is now proposing "automatic Algorithmic Change Protocols" (aACP) and proposals have been advanced to guard against gradual deterioration in prediction quality ("biocreep") (28, 29).

3. Responsible AI in practice

Pillars 1 and 2 should engender well designed and well engineered algorithms, but they do not protect against the unintentional harm that AI may induce. Algorithms can only learn from a digital representation of the world that representation in turn cannot encode moral or ethical standards. Unfair outcomes, discrimination against sub-populations and bias are all reported shortcomings (30). In a dynamic setting, risk can also arise in the form of degraded predictive performance over time. Models that modify clinician's behavior alter patient profiles by design, but predictive success today inevitably erodes future performance by rendering obsolete the historical patterns that drove the performance of the original model (31). Responsible AI in practice requires a systems approach that preempts and safe-guards against these potential risks to patients. We highlight three promising responses to components of this challenge that need to become part of the risk management approach for ML4H.

3.1. Model explainability

We argue that model explainability (Explainable Artificial Intelligence [XAI]) methods need to be prioritized to help systematize and coordinate the processes of model troubleshooting by developers, risk-management by service providers, and system-checks by auditors (32–35). Most AI models that operate as "black-box models" are unsuitable for mission-critical domains, such as healthcare, because they pose risk scenarios where problems that occur can remain

[1] One infamous example from the financial services sector saw a firm lose 170, 000*persecond*(*morethan*400m in 45 min) when an outdated piece of code leaked into production. The firm in question was fined a further <$>12m for "inadequate safeguards" allowing "millions of erroneous orders" (25).

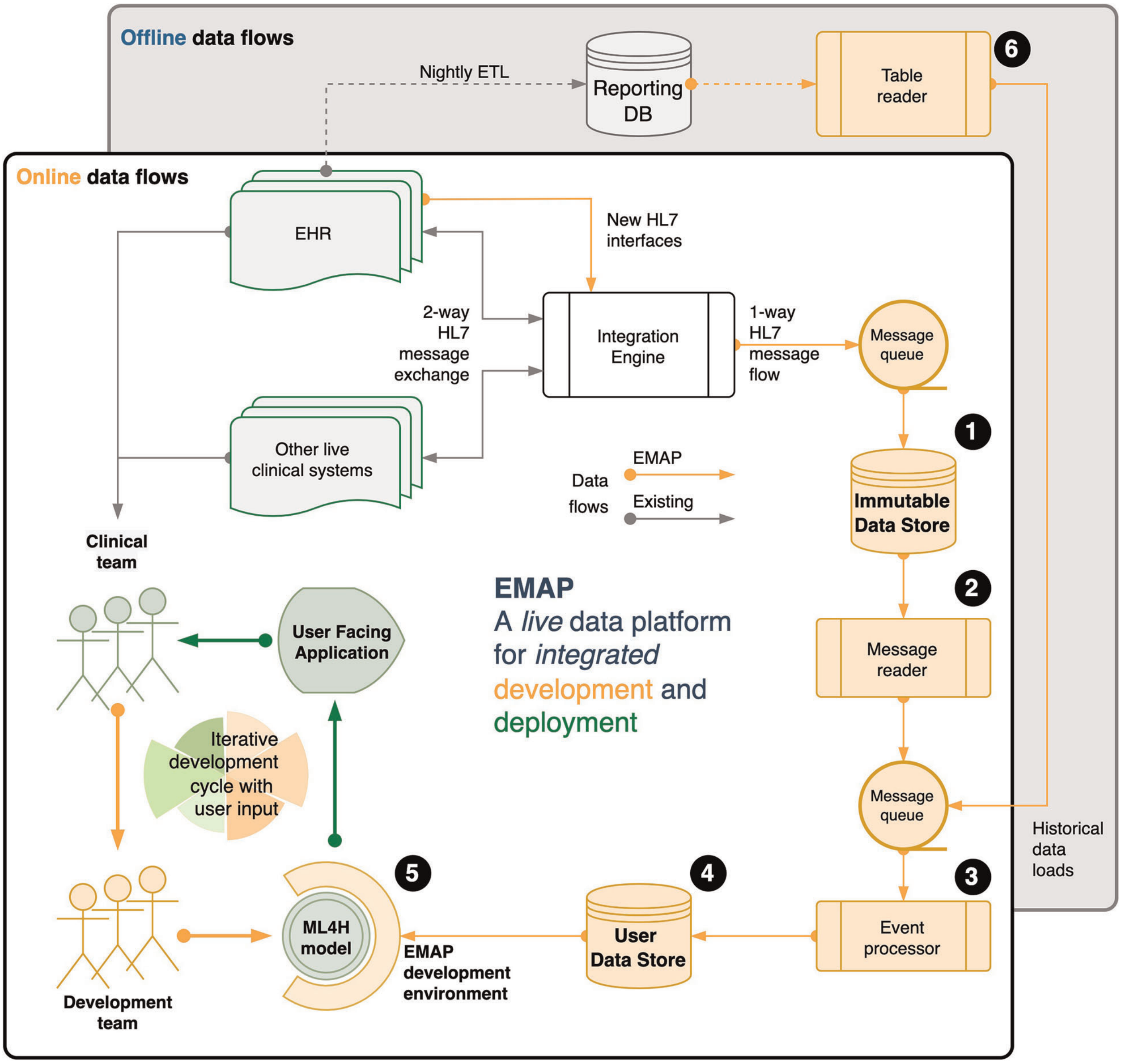


FIGURE 1
Our real-world development is performed on the Experimental Medicine Application Platform (EMAP). EMAP is a clinical laboratory within which ML4H researchers can iteratively build, test and gather feedback from the bedside. It unifies the data and the tools for off-line and online development of ML4H models (see figure and the **(numbers)** in the following sentences that refer to objects in the figure). In brief, EMAP builds a patient orientated SQL database from Health Level 7 version 2 (HL7v2) messages that are being exchanged between hospital systems. HL7v2 messages are ubiquitous in health care, and the de facto standard for internal communication. Rather than multiple pairwise connections between different hospital electronic systems, an integration engine acts as a single hub that routes HL7 messages, and where necessary translates to ensure compatibility. EMAP copies each message passing through the integration engine to a PostgreSQL database, the Immutable Data Store (IDS) **(1)**. A message reader **(2)** processes each live message to an interchange format so that downstream processing is insulated from local HL7 implementation. Separately, the table reader **(6)** processes historical data (e.g. from the reporting database) to the same interchange format. Live messages take priority over historical messages in a queue that feeds the event processor **(3)**. This links each message to a patient and a hospital visit, makes appropriate updates for out of order messages, and merges when separate identifiers are recognised to represent the same patient. A full audit trail is maintained. Each event updates a second live PostgreSQL database, the User Data Store (UDS) **(4)**. The hospital hosts Jupyter and RStudio servers, and a Linux development environment is provided that allows docker deployment, installation of analysis libraries and frameworks, exposes SSH and HTTPS services, and allows user verification against the hospital active directory. **(5)** A typical workflow might include investigation and experimentation in a Jupyter Notebook with data from the UDS, then using a small network of docker containers to run the development script, log outputs to a testing database, and report to users *via* email or a locally hosted web application or dashboard. A fuller explanation is available in the Supplementary Material (Section 2: EMAP data flows).

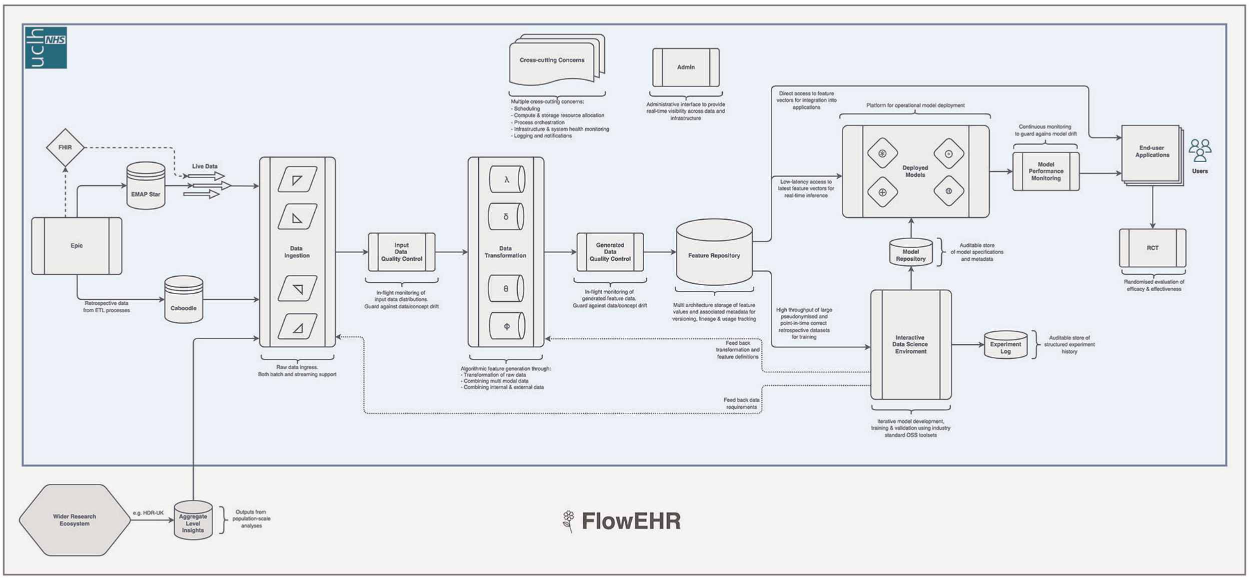

Cross-cutting Concerns
Multiple cross-cutting concerns:
- Scheduling
- Compute & storage resource allocation
- Process orchestration
- Infrastructure & system health monitoring
- Logging and notifications
Admin
Administrative interface to provide real-time visibility across data and infrastructure
FHIR
Live Data
EMAP Star
Epic
Retrospective data from ETL processes
Caboodle
Data Ingestion
Raw data ingress. Both batch and streaming support
Input Data Quality Control
In-flight monitoring of input data distributions. Guard against data/concept drift
λ
δ
θ
φ
Data Transformation
Algorithmic feature generation through:
- Transformation of raw data
- Combining multi modal data
- Combining internal & external data
Generated Data Quality Control
In-flight monitoring of generated feature data. Guard against data/concept drift
Feature Repository
Multi architecture storage of feature values and associated metadata for versioning, lineage & usage tracking
Direct access to feature vectors for integration into applications
Low-latency access to latest feature vectors for real-time inference
High throughput of large pseudonymised and point-in-time correct retrospective datasets for training
Platform for operational model deployment
Deployed Models
Continuous monitoring to guard agains model drift
Model Performance Monitoring
End-user Applications
Users
RCT
Randomised evaluation of efficacy & effectiveness
Model Repository
Auditable store of model specifications and metadata
Interactive Data Science Enviroment
Iterative model development, training & validation using industry standard OSS toolsets
Experiment Log
Auditable store of structured experiment history
Feed back transformation and feature definitions
Feed back data requirements
Wider Research Ecosystem
e.g. HDR-UK
Aggregate Level Insights
Outputs from population-scale analyses
FlowEHR

FIGURE 2
Our ML-Ops platform is called FlowEHR. Moving from left to right across the figure, the system monitors raw input data including checks for distribution shift, builds features with testable and quality controlled code, makes those features available to for both training and predictions to avoid train/serve skew, and maintains an auditable and monitored model repository.

masked and therefore undetectable and unfixable. We acknowledge recent critiques (36, 37) of explainability methods that argue the methods cannot yet be relied on to provide a determinate answer as to whether an AI-recommendation is correct. However, these methods do highlight decision-relevant parts of AI representations, and offer promise in measuring and benchmarking interpretability (38, 39). They are particularly promising for risk management as they can be used to structure a systematic interrogation of the trade-off between interpretability, model accuracy and the risk of model misbehavior.

3.2. Model fail-safes

Prediction models that map patient data to medically meaningful classes are forced to predict without the option to flag users when the model is unsure of an answer. To address this problem, there is good evidence that methods such as Bayesian deep learning and various uncertainty estimates (40) can provide promising ways to detect and refer data samples with high probability of misprediction for human expert review (41–43). These fail safes, or selective prediction approaches should be designed into support systems to preempt and mitigate model misbehavior (29, 44–47). Of note, the European Commission High-Level Expert Group on AI presented guidelines for trustworthy AI in April 2019 with such recommendations: for systems that continue to maintain human-agency *via* a human-in-the-loop oversight. This may even permit less interpretable models to operate when implemented in conjunction with an effective fail-safe system.

3.3. Dynamic model calibration

As discussed, models that influence the evolution of its own future input data are at risk of performance deterioration over time due to input data shifts (48). In such cases, continual learning *via* calibration drift detection and model recalibration (27, 49) provides a promising solution but remains a challenging paradigm in AI. Recalibration with non-stationary incremental data can lead to catastrophic forgetting when the new data negatively interferes with what the model has already learned (50), or a convergence where the model just predicts its own effect and thus should not be updated (31). On the other hand, models can propose poor decisions because of the inherent biases found within the original dataset. In this case, dynamic model recalibration is unlikely to be sufficient and larger model revisions may be required. Here Pillar 1 (RW-dev) with suitable audit and monitoring *via* Pillar 2 (ML-Ops) will be required to overcome what would otherwise be a learning process encumbered by regulatory barriers (51).

4. Implementation science

A well designed, safe, and responsible AI algorithm may still be ineffective if it does not reach a modifiable target on the clinical pathway (19). Unlike medications, algorithms can only effect health by influencing the behavior of clinicians and patients. This translational obstacle parallels the second arm of translational medicine (T2): implementation science (15). Behavior change, in most instances, will be *via* a modification of the choice architecture (passive) (52, 53) or *via* interruptive alerts (active) embedded in the EHR (53). Effective implementation requires a multi-disciplinary approach including human-computer interaction, behavioral science, and qualitative analysis (54).

We strongly argue that this task will be more difficult if done offline and in isolation. Pillar 1 crucially permits not just tuning of the technical performance of the algorithm but rapid build-test-learn cycles that directly involve the target user and the clinical pathway in question. This approach will reduce costs and improve impact, sometimes leading to trade-offs which might appear surprising to those developing away from the bedside (11, 20). This efficiency will again depend on the problem space: where the algorithmic target depends on information arising from the EHR rather than an isolated device or image, and where the pathway involves multiple end-users, then successful implementation will be near impossible if done sequentially (development then deployment) rather than iteratively (54, 55). Academic health science centres must become design "laboratories" where rapid prototyping at the bedside crafts the deployment pathway for *effectiveness* (T2) rather than just efficacy (T1) (15).

Investigations to define how system can influence behavior will need specialist support and tooling. This might require tools embedded within the user interface to evaluate and monitor user interaction, and capture user feedback (56), or directed implementation studies (57).

Despite the oft cited risks of alert fatigue with Clinical Decision Support Systems (CDSS) (58), there is good evidence that well designed alerts can be impactful (53, 59, 60). Overt behavioural modifications will need a mechanism to explain their recommendation (as per XAI) or generate trust (see Pillar 5) (61). Trust will possibly be more important where behavior modification is indirect through non-interruptive techniques (e.g. re-ordering preference lists or otherwise adapting the user interface to make the recommended choice more accessible).

5. Continuous clinical evaluation

Our analogy with translational medicine breaks down at the evaluation stage. For drug discovery, evaluation is *via* a randomized controlled trial (RCT). Randomization handles unanticipated bias and ML4H should hold itself to the same standard but of 350,000 studies registered on ClinicalTrials.gov in 2020, just 358 evaluated ML4H, and only 66 were randomized (62). As usual for ML4H, those RCTs were not interacting with EHR data. They were evaluations of algorithms supporting imaging, cataract screening, colonoscopy, cardiotocographs and more (63–69).

Where the ML4H intervention delivers a novel biological treatment strategy, then it is appropriate to reach for the full paraphenalia used in Clinical Trials of Investigational Medicinal Products (CTIMPs) (2). But in many cases, algorithms will be used to optimize operational workflows and clinical pathways. These pathways may be specific and contextual rather than generalizable. Poor external validity is not a critique: an algorithm that is useful or important in one institution does not have to be relevant in the next (the "myth of generalizability") (70). Moreover, the algorithm is not the same as the patented and fixed active ingredient in a medicinal product. This is no single point in time nor single host environment at which it can be declared enduringly effective. This means that institutions deploying and relying on these tools need a strategy for rapid continuous clinical and operational evaluation.

This time the EHR may provide an advantage instead of just additional complexity. Since ML4H algorithms must be implemented through some form of direct or indirect CDSS, then the next logical step is to randomize the deployment of those alerts. This in itself is not novel. Randomized deterministic alerts from CDSS are part of the standard evaluation toolkit for quality improvement initiatives in at NYU Langone (71), and for research elsewhere (72). At NYU Langone, such tooling permitted a small team to deliver 10 randomized trials within a single year (71).

The final pillar in our CDE uses the same approach for the probabilistic insights derived from ML4H. Excellent patient and public involvement, and ethical guidance, will be required to distinguish those algorithms that require per patient point-of-care consent from those that can use opt-out or cluster methods. But we think that latter group is large for two reasons. Firstly, patients are exposed to varying treatment regimes by dint of their random interaction with different clinicians based on geography (the healthcare provider they access) and time (staff holidays and shift patterns etc.). This routine variation in practice is summarized as the 60-30-10 problem: 60% of care follows best practice; 30% is wasteful or ineffective and 10% is harmful (6). Secondly, because the intervention is informational, there is ethical precedent for patient level randomization without consent (e.g. Acute Kidney Injury alerts) (72). This hints at a larger and more routine role for randomization in evaluation of algorithms. This in turn is supported by a growing (52, 73, 74) but sometimes conflicting (75) literature on opt-out consent in

Learning Healthcare Systems (LHS). As such, progress will require careful attention to a range of concerns.

At our own institution, we have extended this ethical and safety case one step further, and we are piloting a study design where the randomization is non-mandatory: a nudge not an order (76). The clinician is explicitly invited to only comply with the randomization where they have equipoise themselves. Where they have a preference, they overrule the alert (see Vignette 1 in the **Supplementary Material**).

Embedded randomized digital evaluation should permit rapid evidence generation, and build the trust needed to support the implementation described under Pillar 4.

Drug discovery parallels

We have described a template for a Clinical Deployment Environment that supports the translation of ML4H algorithms from bench to bedside. Although the requirements differ, the objective is similar to that for drug development. A similar approach to phasing has previously been proposed for (biomarker) prediction models (77).

Most ML4H that derives value from the EHR is in the pre-clinical phase. In drug development, the objective of this phase is to identify candidate molecules which might make effective drugs. Evaluation is conducted *in vitro*. Metrics used to evaluate candidates, such binding affinity or other pharmacokinetic properties, describe the properties of the molecule (78). For ML, the objective is to identify candidate algorithms, comprising of input variables and model structures, which might make the core of an effective CDSS. Evaluation is conducted offline on de-identified datasets. Metrics used to evaluate candidates, such Area Under the Receiver Operator Curve (AUROC), the F1 score and calibration, describe the properties of the algorithm (79).

Phase 1 drug trials are the first time a drug candidate is tested in humans. They are conducted in small numbers of healthy volunteers. The aim of the trial is to determine the feasibility of progressing to trials in patients by determining drug safety and appropriate dosage. Drug formulation, the processes by which substances are combined with the active pharmaceutical ingredient to optimize the acceptability and effective delivery of the drug, is also considered at this stage. Phase 1 ML4H trials are the first time an algorithm candidate is tested within the healthcare environment. The aim of the trial is to determine the feasibility of progressing to trials of efficacy by ensuring the algorithm implementation is safe, reliable and able to cope with real-world data quality issues. The development of a mechanism to deliver of algorithm outputs embedded in the clinical workflow is also be considered at this stage.

Phase 2 drug trials involve recruitment of small numbers patients with the disease of interest, typically 50–200. The aim is to determine drug efficacy at treating the disease. Treating clinicians are involved in so far as they must agree to prescribe the drug for their patients. The trials are often too short to determine long term outcomes, therefore surrogate measures such biomarker status or change in tumour size are used as endpoints (80). Phase 2 ML4H trials involve recruitment of small numbers of clinicians making the decision of interest, typically 5–10. The aim is to determine the efficacy of the algorithm in improving their decisions. Patients are involved in so far as they must agree to be on the receiving end of these supported decisions and identifiable data is required. Endpoints are markers of successful task completion in all cases. Investigations to determine ways in which the system could be more successful in influencing user behavior are carried out at this stage. These include usability analyses, considerations of how well the ML4H/CDSS is integrated into the overall system and implementation studies to identify how best to optimize end-user adoption and engagement (57).

Phase 3 drug trials involve the recruitment of large numbers of patients to determine whether a drug is effective in improving patient outcomes. The gold standard of trial design is a double-blinded randomized controlled trial (RCT). Phase 3 ML4H trials will require integration of data from multiple centers for algorithms acting on specific decisions but inevitably adapted to their local data environment.

The phases of drug development are not meant to be matched 1:1 to the pillars of the CDE described here: in fact, our argument for "real-world" *development* deliberately seeks to merge the steps. But the parallel is drawn to highlight the effort necessary to see ML4H have an impact on the clinical and operational decision making in the workplace. Heretofore this effort has been hugely underestimated.

Conclusion

Even this analogy stops short of the full task of deployment. With drug development, the universities and the pharmaceutical industry go on to take advantage of a supply chain to deliver the drug to the hospital with the necessary quality control and monitoring. Those prescribing and administering the drug have spent years in training, and are supported by pharmacists and medication safety experts. And even after the drug is administered, observation and long term follow-up continue to identify side-effects and long term hazards.

That network of expertise and infrastructure is largely in place where software *is within* (not *as*) a medical device, but is only just being envisioned where the data driving ML4H comes from the EHR. This distinction needs to be made else the disillusionment with the promise of ML4H will continue. The technology does have the potential to change how we deliver health but the methodology alone is insufficient. The impressive demonstrations of the power of AI and ML to beat humans in games, and predict protein structures does not mean that these tools are ready for wide spread deployment.

But we should not be pessimistic. As per author William Gibson, it is clear that "The Future Has Arrived — It"s Just Not Evenly Distributed." Beyond healthcare, machine learning has already demonstrated that it can reliably create value (5). It is now our responsibility to take those lessons and adapt them for our patients.

The Five Pillars outlined here are a sketch of that redistribution. They are born from our local experience (Pillars 1, 2 and 5) and our wider observations (Pillars 3 and 4). They fundamentally are an argument for a professionalization of ML4H, and a caution against the "get-rich quick" headlines in the popular and scientific press (1). We envision a future where each algorithm is managed in a digital pharmacy with the same rigor that we apply to medicines. But unlike drugs, some of these algorithms will have their entire life-cycle, from development to deployment, managed by the local healthcare provider. Computer vision tasks that support diagnostic radiology can be partially developed offline. Components of sepsis prediction tools will transfer from institution to institution but will need adapting to local clinical workflows. But there will be opportunity and value for ML4H to optimize operational tasks that are temporary or specific to that institution. This means that some development and much of the deployment will require a suitably trained workforce, and an infrastructure perhaps supported by these five pillars.

Author contributions

All authors listed have made a substantial, direct and intellectual contribution to the work, and approved it for publication. All authors contributed to the article and approved the submitted version.

Funding

SH is supported by a Health Foundation Improvement Science Fellowship, and by funding from the Biomedical Research Centre based at the University College Hospitals (UCLH) National Health Service (NHS) Foundation Trust and University College London (UCL). NS is supported by funding from Health Data Research UK. SH, TK, and NS are supported funds from the National Institute for Health Research (Artificial Intelligence, Digitally adapted, hyper-local real-time bed forecasting to manage flow for NHS wards, AI AWARD01786) and NHS-X. WL is supported by a UKRI Ernest Rutherford Fellowship. The views expressed in this publication are those of the author(s) and not necessarily those of the National Institute for Health Research, NHS-X or the Department of Health and Social Care.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2022.939292/full#supplementary-material.

References

1. Bunz M, Braghieri M. The AI doctor will see you now: Assessing the framing of AI in news coverage. *AI Soc.* (2022) 37:9–22. doi: 10.1007/s00146-021-01145-9

2. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med.* (2018) 24:1716–20. doi: 10.1038/s41591-018-0213-5

3. Hyland SL, Faltys M, Hüser M, Lyu X, Gumbsch T, Esteban C, et al. Early prediction of circulatory failure in the intensive care unit using machine learning. *Nat Med.* (2020) 26:364–73. doi: 10.1038/s41591-020-0789-4

4. Tomašev N, Glorot X, Rae J, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. *Nature.* (2019) 572:116–9. doi: 10.1038/s41586-019-1390-1

5. McRae H. *Facebook, Airbnb, Uber, and the unstoppable rise of the content non-generators.* London: The Independent (2015).

6. Braithwaite J, Glasziou P, Westbrook J. The three numbers you need to know about healthcare: The 60-30-10 challenge. *BMC Med.* (2020) 18:1–8. doi: 10.1186/s12916-019-1443-1

7. The national strategy for AI in health and social care (2022). Available at: https://www.nhsx.nhs.uk/ai-lab/ai-lab-programmes/the-national-strategy-for-ai-in-health-and-social-care/ (Accessed July 1, 2022).

8. Digital future index 2021–2022 (2021). Available at: https://www.digicatapult.org.uk/news-and-insights/publications/post/digital-future-index-2021-2022/ (Accessed July 1, 2022).

9. Everson J, Rubin JC, Friedman CP. Reconsidering hospital EHR adoption at the Dawn of HITECH: Implications of the reported 9% adoption of a "basic" EHR. *J Am Med Inform Assoc*. (2020) 27:1198–205. doi: 10.1093/jamia/ocaa090

10. Muehlematter U, Daniore P, Vokinger K. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015-20): A comparative analysis. *Lancet Digit Health*. (2021) 3:e195–203. doi: 10.1016/S2589-7500(20)30292-2

11. Morse K, Bagley S, Shah N. Estimate the hidden deployment cost of predictive models to improve patient care. *Nat Med*. (2020) 26:18–9. doi: 10.1038/s41591-019-0651-8

12. Eaneff S, Obermeyer Z, Butte AJ. The case for algorithmic stewardship for artificial intelligence and machine learning technologies. *JAMA*. (2020) 324:1397–8. doi: 10.1001/jama.2020.9371

13. Burton PR, Murtagh MJ, Boyd A, Williams JB, Dove ES, Wallace SE, et al. Data safe havens in health research and healthcare. *Bioinformatics*. (2015) 31:3241–8. doi: 10.1093/bioinformatics/btv279

14. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. *npj Digit Med*. (2018) 1:40. doi: 10.1038/s41746-018-0048-y

15. Woolf SH. The meaning of translational research and why it matters. *JAMA*. (2008) 299:211–3. doi: 10.1001/jama.2007.26

16. Corrigan-Curay J, Sacks L, Woodcock J. Real-World evidence and real-world data for evaluating drug safety and effectiveness. *JAMA*. (2018) 320:867. doi: 10.1001/jama.2018.10136.

17. Williamson E, Walker A, Bhaskaran K, Bacon S, Bates C, Morton C, et al. Factors associated with COVID-19-related death using OpenSAFELY. *Nature*. (2020) 584:430–6. doi: 10.1038/s41586-020-2521-4

18. Data research infrastructure landscape: A review of the UK data research infrastructure (2021). Available at: https://dareuk.org.uk/wp-content/uploads/2021/11/DARE_UK_Data_Research_Infrastructure_Landscape_Review_Oct_2021.pdf (Accessed July 1, 2022).

19. The DECIDE-AI Steering Group. DECIDE-AI: New reporting guidelines to bridge the development-to-implementation gap in clinical artificial intelligence. *Nat Med*. (2021) 27:186–7. doi: 10.1038/s41591-021-01229-5

20. Shah NH, Milstein A, Bagley SC. Making machine learning models clinically useful. *JAMA*. (2019) 322:1351. doi: 10.1001/jama.2019.10306.

21. Guinney J, Saez-Rodriguez J. Alternative models for sharing confidential biomedical data. *Nat Biotechnol*. (2018) 36:391–2. doi: 10.1038/nbt.4128

22. DevOps (2022). Available at: https://en.wikipedia.org/wiki/DevOps (Accessed July 1, 2022).

23. Ben-Israel D, Jacobs WB, Casha S, Lang S, Ryu WHA, de Lotbiniere-Bassett M, et al. The impact of machine learning on patient care: A systematic review. *Artif Intell Med*. (2020) 103:101785. doi: 10.1016/j.artmed.2019.101785

24. Sculley D, Holt G, Golovin D, Davydov E, Phillips T, Ebner D, et al. Hidden technical debt in machine learning systems. *Adv Neural Inf Process Syst*. (2015) 28:2503–11. Available at: https://dl.acm.org/doi/10.5555/2969442.2969519

25. SEC Charges Knight Capital With Violations of Market Access Rule (2013). Available at: https://www.sec.gov/news/press-release/2013-222 (Accessed April 27, 2022).

26. John MM, Olsson HH, Bosch J. Towards MLOps: A framework and maturity model, In: John MM, Olsson HH, Bosch J, editors. *Towards MLOps: A Framework and Maturity Model.* 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA) (2021). p. 1–8. doi: 10.1109/SEAA53835.2021.00050

27. Feng J, Phillips RV, Malenica I, Bishara A, Hubbard AE, Celi LA, et al. Clinical artificial intelligence quality improvement: Towards continual monitoring and updating of AI algorithms in healthcare. *npj Digit Med*. (2022) 5:66. doi: 10.1038/s41746-022-00611-y

28. Artificial Intelligence and Machine Learning (AI/ML) Software as a Medical Device Action Plan. US Food & Drug Administration (2021). Available at: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device?mc_cid=20dc2074ab&mc_eid=c49edc17d2 (Accessed July 1, 2022).

29. Feng J, Emerson S, Simon N. Approval policies for modifications to machine learning-based software as a medical device: A study of bio-creep. *Biometrics*. (2021) 77:31–44. doi: 10.1111/biom.13379.

30. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in AI safety. arXiv preprint, arXiv:160606565 (2016). doi: 10.48550/arXiv.1606.06565

31. Liley J, Emerson S, Mateen B, Vallejos C, Aslett L, Vollmer S. Model updating after interventions paradoxically introduces bias. *Proc Mach Learn Res*. (2021) 130:3916–24. doi: 10.48550/arXiv.2010.11530. Available at: http://proceedings.mlr.press/v130/liley21a.html

32. Gunning D, Stefik M, Choi J, Stumpf S, Yang G. XAI—explainable artificial intelligence. *Sci Robot*. (2019) 4:1–2. doi: 10.1126/scirobotics.aay7120

33. Mueller S, Hoffman R, Clancey W, Emrey A, Klein G. Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI. arXiv preprint, arXiv:190201876 (2019).

34. Vilone G, Longo L. Explainable artificial intelligence: A systematic review. arXiv preprint, arXiv:200600093 (2020).

35. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A review of machine learning interpretability methods. *Entropy*. (2020) 23:1–45. doi: 10.3390/e23010018

36. Ghassemi M, Oakden-Rayner L, Beam A. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health*. (2021) 3:e745–50. doi: 10.1016/S2589-7500(21)00208-9

37. The Disagreement Problem in Explainable Machine Learning: A Practitioner's Perspective (2022). Available at: http://arxiv.org/abs/2202.01602 (Accessed June 29, 2022).

38. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv preprint, arXiv:170208608 (2017).

39. Hoffman R, Mueller S, Klein G, Litman J. Metrics for explainable AI: Challenges and prospects. *arXiv preprint*, arXiv:181204608 (2018).

40. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf Fusion*. (2021) 76:243–97. doi: 10.1016/j.inffus.2021.05.008

41. Leibig C, Allken V, Ayhan M, Berens P, Wahl S. Leveraging uncertainty information from deep neural networks for disease detection. *Sci Rep*. (2017) 7:1–14. doi: 10.1038/s41598-017-17876-z

42. Filos A, Farquhar S, Gomez A, Rudner T, Kenton Z, Smith L, et al. A systematic comparison of Bayesian deep learning robustness in diabetic retinopathy tasks. arXiv preprint, arXiv:191210481 (2019).

43. Ghoshal B, Tucker T. Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. arXiv preprint, arXiv:200310769 (2020).

44. Chow C. On optimum recognition error and reject tradeoff. *IEEE Trans Inf Theory*. (1970) 16:41–6. doi: 10.1109/TIT.1970.1054406

45. Bartlett PL, Wegkamp MH. Classification with a reject option using a hinge loss. *J Mach Learn Res* (2008) 9:18. Available at: http://jmlr.org/papers/v9/bartlett08a.html

46. Tortorella F. An optimal reject rule for binary classifiers. In: G Goos, J Hartmanis, J van Leeuwen, FJ Ferri, JM Iñesta, A Amin, P Pudil, editors. *Advances in pattern recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg (2000). p. 611–20. doi: 10.1007/3-540-44522-6_63

47. El-Yaniv R, Wiener Y. On the foundations of noise-free selective classification. *J Mach Learn Res*. (2010) 11:37. doi: 10.5555/1756006.1859904

48. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc*. (2017) 24:1052–61. doi: 10.1093/jamia/ocx030

49. Davis SE, Greevy RA, Lasko TA, Walsh CG, Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. *J Biomed Inform*. (2020) 112:103611. doi: 10.1016/j.jbi.2020.103611

50. Parisi G, Kemker R, Part J, Kanan C, Wermter S. Continual lifelong learning with neural networks: A review. *Neural Netw*. (2019) 113:54–71. doi: 10.1016/j.neunet.2019.01.012

51. Lee CS, Lee AY. Clinical applications of continual learning machine learning. *Lancet Digit Health*. (2020) 2:e279–81. doi: 10.1016/S2589-7500(20)30102-3

52. Halpern SD. Using default options and other nudges to improve critical care. *Crit Care Med*. (2018) 46:460–4. doi: 10.1097/CCM.0000000000002898

53. Main C, Moxham T, Wyatt JC, Kay J, Anderson R, Stein K. Computerised decision support systems in order communication for diagnostic, screening or monitoring test ordering: Systematic reviews of the effects and cost-effectiveness of systems. *Health Technol Assess*. (2010) 14:1–227. doi: 10.3310/hta14480.

54. Sendak M, Ratliff W, Sarro D, Alderton E, Futoma J, Gao M, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. *JMIR Med Inform*. (2020) 8:e15182. doi: 10.2196/15182

55. Connell A, Black G, Montgomery H, Martin P, Nightingale C, King D, et al. Implementation of a digitally enabled care pathway (part 2): Qualitative analysis

of experiences of health care professionals. *J Med Internet Res*. (2019) 21:e13143. doi: 10.2196/13143

56. Yusop NSM, Grundy J, Vasa R. Reporting usability defects: A systematic literature review. *IEEE Trans Softw Eng*. (2017) 43:848–67. doi: 10.1109/TSE.2016.2638427.

57. Sutton RT, Pincock D, Baumgart DC, Sadowski DC, Fedorak RN, Kroeker KI. An overview of clinical decision support systems: Benefits, risks, and strategies for success. *npj Digit Med*. (2020) 3:1–10. doi: 10.1038/s41746-020-0221-y

58. Phansalkar S, van der Sijs H, Tucker A, Desai A, Bell D, Teich J, et al. Drug-drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records. *J Am Med Inform Assoc*. (2013) 20:489–93. doi: 10.1136/amiajnl-2012-001089

59. Park G, Kang B, Kim S, Lee J. Retrospective review of missed cancer detection and its mammography findings with artificial-intelligence-based, computer-aided diagnosis. *Diagnostics*. (2022) 12:387. doi: 10.3390/diagnostics12020387

60. Sayres R, Taly A, Rahimy E, Blumer K, Coz D, Hammel N, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology*. (2019) 126:552–64. doi: 10.1016/j.ophtha.2018.11.016

61. McCoy L, Brenna C, Chen S, Vold K, Das S. Believing in black boxes: Machine learning for healthcare does not need explainability to be evidence-based. *J Clin Epidemiol*. (2022) 142:252–7. doi: 10.1016/j.jclinepi.2021.11.001

62. Zippel C, Bohnet-Joschko S. Rise of clinical studies in the field of machine learning: A review of data registered in ClinicalTrials.gov. *Int J Environ Res Public Health*. (2021) 18:5072. doi: 10.3390/ijerph18105072

63. CG INFANT. Computerised interpretation of fetal heart rate during labour (INFANT): A randomised controlled trial. *Lancet*. (2017) 389:1719–29. doi: 10.1016/S0140-6736(17)30568-8

64. Titano JJ, Badgeley M, Schefflein J, Pain M, Su A, Cai M, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. *Nat Med*. (2018) 24:1337–41. doi: 10.1038/s41591-018-0147-y

65. Wang P, Berzin T, Glissen Brown J, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: A prospective randomised controlled study. *Gut*. (2019) 68:1813–9. doi: 10.1136/gutjnl-2018-317500

66. Wu L, Zhang J, Zhou W, An P, Shen L, Liu J, et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. *Gut*. (2019) 68:2161–9. doi: 10.1136/gutjnl-2018-317366

67. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: A multicentre randomized controlled trial. *EClinicalMedicine*. (2019) 9:52–9. doi: 10.1016/j.eclinm.2019.03.001

68. Turakhia MP, Desai M, Hedlin H, Rajmane A, Talati N, Ferris T, et al. Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The apple heart study. *Am Heart J*. (2019) 207:66–75. doi: 10.1016/j.ahj.2018.09.002

69. Long E, Lin H, Liu Z, Wu X, Wang L, Jiang J, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. *Nat Biomed Eng*. (2017) 1:1–8. doi: 10.1038/s41551-016-0024

70. Futoma J, Simons M, Panch T, Doshi-Velez F, Celi LA. The myth of generalisability in clinical research and machine learning in health care. *Lancet Digit Health*. (2020) 2:e489–92. doi: 10.1016/S2589-7500(20)30186-2

71. Horwitz LI, Kuznetsova M, Jones SA. Creating a learning health system through rapid-cycle, randomized testing. *N Engl J Med*. (2019) 381:1175–9. doi: 10.1056/NEJMsb1900856

72. Wilson FP, Martin M, Yamamoto Y, Partridge C, Moreira E, Arora T, et al. Electronic health record alerts for acute kidney injury: Multicenter, randomized clinical trial. *Br Med J*. (2021) 372:m4786. doi: 10.1136/bmj.m4786

73. London AJ. Learning health systems, clinical equipoise and the ethics of response adaptive randomisation. *J Med Ethics*. (2018) 44:409–15. doi: 10.1136/medethics-2017-104549

74. Scobie S, Castle-Clarke S. Implementing learning health systems in the UK NHS: Policy actions to improve collaboration and transparency and support innovation and better use of analytics. *Learn Health Syst*. (2020) 4:e10209. doi: 10.1002/lrh2.10209

75. Meyer MN, Heck PR, Holtzman GS, Anderson SM, Cai W, Watts DJ, et al. Objecting to experiments that compare two unobjectionable policies or treatments. *Proc Natl Acad Sci USA*. (2019) 116:10723–8. doi: 10.1073/pnas.1820701116

76. Wilson MG, Asselbergs FW, Harris SK. Learning from individualised variation for evidence generation within a learning health system. *Br J Anaesth*. (2022) 128:e320–2. doi: 10.1016/j.bja.2022.02.008

77. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, et al. Phases of biomarker development for early detection of cancer. *J Natl Cancer Inst*. (2001) 93:1054–61. doi: 10.1093/jnci/93.14.1054

78. Lo Y-C, Rensi SE, Torng W, Altman RB. Machine learning in chemoinformatics and drug discovery. *Drug Discov Today*. (2018) 23:1538–46. doi: 10.1016/j.drudis.2018.05.010

79. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Ann Intern Med*. (2015) 162:55–63. doi: 10.7326/M14-0697

80. Van Norman GA. Phase II trials in drug development and adaptive trial design. *JACC: Basic Transl Sci*. (2019) 4:428–37. doi: 10.1016/j.jacbts.2019.02.005

TYPE Perspective
PUBLISHED 24 August 2022
DOI 10.3389/fdgth.2022.931439

OPEN ACCESS

EDITED BY
Mark Sendak,
Other, United States

REVIEWED BY
Alessandro Blasimme,
ETH Zürich, Switzerland
Laurie Lovett Novak,
Vanderbilt University Medical Center, United States

*CORRESPONDENCE
Frank Liao
fjliao@wisc.edu

SPECIALTY SECTION
This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health

RECEIVED 29 April 2022
ACCEPTED 18 July 2022
PUBLISHED 24 August 2022

CITATION
Liao F, Adelaine S, Afshar M and Patterson BW (2022) Governance of Clinical AI applications to facilitate safe and equitable deployment in a large health system: Key elements and early successes.
Front. Digit. Health 4:931439.
doi: 10.3389/fdgth.2022.931439

Governance of Clinical AI applications to facilitate safe and equitable deployment in a large health system: Key elements and early successes

Frank Liao[1,2]*, Sabrina Adelaine[2], Majid Afshar[3,4] and Brian W. Patterson[1,2,4,5]

[1]BerbeeWalsh Department of Emergency Medicine, UW-Madison, Madison, WI, United States, [2]Department of Information Services, UW Health, Madison, WI, United States, [3]Department of Medicine, UW-Madison, Madison, WI, United States, [4]Department of Biostatistics and Medical Informatics, UW-Madison, Madison, WI, United States, [5]Department of Industrial and Systems Engineering, UW-Madison, Madison, WI, United States

One of the key challenges in successful deployment and meaningful adoption of AI in healthcare is health system-level governance of AI applications. Such governance is critical not only for patient safety and accountability by a health system, but to foster clinician trust to improve adoption and facilitate meaningful health outcomes. In this case study, we describe the development of such a governance structure at University of Wisconsin Health (UWH) that provides oversight of AI applications from assessment of validity and user acceptability through safe deployment with continuous monitoring for effectiveness. Our structure leverages a multi-disciplinary steering committee along with project specific sub-committees. Members of the committee formulate a multi-stakeholder perspective spanning informatics, data science, clinical operations, ethics, and equity. Our structure includes guiding principles that provide tangible parameters for endorsement of both initial deployment and ongoing usage of AI applications. The committee is tasked with ensuring principles of interpretability, accuracy, and fairness across all applications. To operationalize these principles, we provide a value stream to apply the principles of AI governance at different stages of clinical implementation. This structure has enabled effective clinical adoption of AI applications. Effective governance has provided several outcomes: (1) a clear and institutional structure for oversight and endorsement; (2) a path towards successful deployment that encompasses technologic, clinical, and operational, considerations; (3) a process for ongoing monitoring to ensure the solution remains acceptable as clinical practice and disease prevalence evolve; (4) incorporation of guidelines for the ethical and equitable use of AI applications.

KEYWORDS

Clinical AI, AI, AI adoption, predictive analytics, governance, oversight, ethics, equity

Introduction

Artificial intelligence (AI) holds the promise to transform clinical care (1), and is increasingly being used in clinical practice. However, appropriate governance of these models remains in its infancy, especially as larger governing bodies like the Food and Drug Administration and World Health Organization are trying to keep up with the advancements in technology and its role in health care. Unlike other sectors, healthcare must carry a lower tolerance for error and bias as AI-driven tools have a direct impact on patient lives and unchecked errors may cause harm or death (2). UW Health, like many academic institutions, frequently encounters new commercial products and scientific innovations that leverage AI for healthcare delivery in both diagnostics and prognosis. While several groups have discussed the importance of, and methodologies for, responsible development of these interventions (3), the accountability of safe and effective deployment of AI-driven applications ultimately falls onto the health system. As the ethics surrounding AI-development has received increasing scrutiny (4), there has been little literature focusing on institutional governance. As we expand our technical ability to provide solutions, more skepticism and questions surface, and at times resistance, around the suitability of using AI in routine clinical care from all levels of the organization, ranging from front-line clinical staff to executive leadership. In response to these questions and the challenges for implementation, the health system recognized the need for a governance structure to endorse and oversee adoption, implementation, and ongoing value evaluation of AI-driven applications. This case study describes the development and nature of governance of clinical AI applications at our institution.

The role of governance

Challenges

During deployment of the first set of AI-driven applications, we encountered several challenges unique to the field. From a systems perspective these challenges can be grouped into three domains, where each domain represents a particular constituency with associated considerations (**Table 1**). The first domain is clinical, and its constituents are the patients, clinicians and other front-line users of the AI solution. The challenges associated with the first domain are related to clinical acceptability of the AI output, and actionability (in terms of personal agency), as well as explainability. The goal of governance for this domain is maintaining patient safety, as well as securing clinician acceptance and adoption.

TABLE 1 Challenges to the adoption of AI categorized by domains with associated constituencies and goals.

Domain	Constituents	Goals of governance
Clinical	Patients, clinicians, staff	Patient safety, model effectiveness, explainability and adoption
Operational	Clinical and operational stakeholders	Integration of AI models into routine health system operations
Leadership	Hospital and health system leaders	Endorsement by senior leadership, integration into overall strategic plans

The second domain is operational; the constituents of this domain represent the systems-level components that are part the care delivery mechanism. This group includes the stakeholders that represent clinical operations, information services and informatics. The challenges for the second domain are related to actionability (in terms of clinical protocols and governance), performance validity, sustainability, and accountability (in terms of ongoing support mechanisms). The goal of the governance for this domain is complementary oversight that is compatible with the routine operating model of the health system.

The third domain is leadership and its constituents are those who manage the strategic direction of the health system, hold key decision rights, and govern the resources. The challenges for the third domain are related to oversight, accountability, and equitability. The goal of the governance for this domain is endorsement by senior leadership in health operations.

As we operationalized predictive models, we surfaced challenges in each domain. Some domains, such as clinical and operational, required more focus earlier in development, with rapid adaption and evolution, while the leadership domain, adapted at a different rate that required more focus with commensurate experience of the organization. The focus of the domain shifted and adapted over time depending on the maturity of each domain and the individualized needs of each AI application.

To address the challenges in the first and second domain during rollout of our initial models, individual solution workgroups were established in an *ad hoc* fashion. The responsibility of these workgroups was performing due diligence and providing detailed scrutiny of the AI solutions to establish the necessary validity both clinically, technically, and ethically. Examples of specific activities include retrospective and prospective validations of the performance of the AI solution on the UWH patient population as a whole and on specific demographic sub-populations; closer examination of the clinical inputs or variables used by the AI solution; the suitability of the solution's output within its specific operational and clinical context; and ongoing assessment of the solution's performance, clinician adoption and usage, and other related metrics.

The output of the workgroups was also synthesized and disseminated to the constituents in the second domain to procure complementary governance such as the approval of new or updated clinical protocols that incorporated the AI solution, or operational buy-in to update the ongoing or routine processes of the health system. The designation of these workgroups evolved along with the AI maturity of the organization, beginning with "algorithm workgroups", then to "algorithm committees", with a current designation of "algorithm sub-committees". Initially, challenges within the third domain, leadership, were addressed through executive sponsorship of a specific use case, which constituted simple endorsement for smaller applications and executive steering committees when necessary for larger projects.

The composition and membership of workgroups were multi-disciplinary, as necessary to perform their function. Key disciplines included clinical subject matter expertise, data science, informatics, information technology, clinical operations, bioethics, human factors or design thinking. Common roles that were represented included physicians, nurses, data scientists, analytics professionals, information services, clinical quality, and academic faculty. The primary advantage of these workgroups was the ability to bring the content and methodologic expertise a solution required for operationalization. Furthermore, by combining clinical and operational considerations for narrowly focused use cases, these workgroups were able to maintain a nimble, innovative approach to each use case.

However, after several solutions were enacted, weaknesses of these *ad hoc* workgroups in addressing challenges from operational and leadership domains became more apparent. From a system-wide standpoint, more integration was needed to create visibility and oversight of all models, and retain consistent governance across a variety of clinical use cases. To address these challenges while keeping the advantages of the individual use case workgroups, we created an institutional-level steering committee which would provide a front door and maintain oversight of all models while retaining individual workgroups for more detailed governance. This "Clinical AI and Predictive Analytics Committee" is multi-disciplinary and included a superset of the same disciplines that comprised the use-case specific algorithm workgroups. Figures 1, 2 show the composition and representation of the institutional committee and its relationship with the sub-committees, respectively. However, one advantage of an institutional-level committee was a stronger ethics and equity perspective. Another advantage of institution-level committee is a clear and strong connection to the University of Wisconsin campus. Connection to the campus brings academic expertise in the relevant domains, and the associated research enterprise including complementary guidelines to other institutions like the Institutional Review Board (IRB). The committee functions as a front door for the evaluation and vetting of predictive solutions prior to implementation, and for new models it commissions and oversees workgroups. The committee reports up to existing clinical and informatics leadership structures in the university and health system and provides visibility on all clinical uses of AI to these groups.

The institutional-level committee defines and establishes definitions of key terms such as "predictive model" as well as guiding principles. However, given the broad scope of potential AI applications, the committee does not seek to perform all of the duties previously performed by algorithm workgroups for each application. Rather, once an application is brought to the committee, it commissions an "Algorithm sub-committee" with the scope of a specific application. Each sub-committee follows the established guiding principles and applies them when evaluating the algorithm(s) for its specific use case, and reports back to the institutional Clinical AI and predictive Analytics Committee. This federated system seeks to retain the benefits of the application specific workgroups while realizing the advantages of a single committee to govern all applications.

Scope

The committee oversees AI and predictive solutions which affect clinical care in the health system, including workflow and implementation. This includes solutions aimed at clinical care (e.g. patient deterioration or sepsis), patient access and resource allocation (e.g. length-of-stay (LOS) predictions, inpatient capacity management). The committee does not oversee models in which there is no clear connection to clinical care (e.g., a financial model to predict likelihood of payment). AI as a component of an FDA approved medical device is not necessarily overseen if a model isn't modifiable at the health system level and its performance has been well-characterized (e.g., an FDA approved software program to evaluate diabetic retinopathy from retinal pictures).

Guiding principles

Below are the currently endorsed set of guiding principles:

1. Predictive model (including outside vendors or internal innovation) evaluation includes validation of performance on UW Health production data and clinician review against the appropriate target labels for application.
2. Model evaluation includes statistical measures (e.g., sensitivity, specificity, PPV) and relevant operational and health metrics (e.g., alarm rate, work-up to detection ratio, appropriate use, fairness, cost-effectiveness and intervention effectiveness on health outcomes).

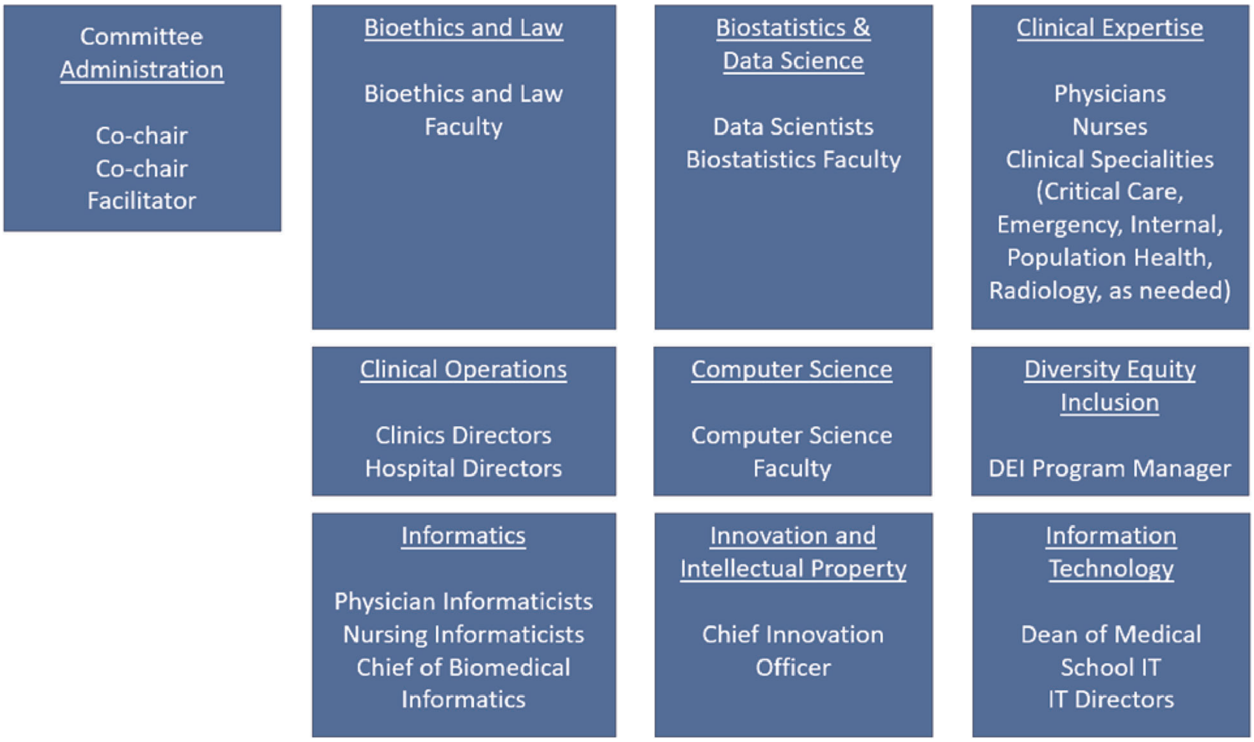


FIGURE 1
Clinical AI and predictive analytics committee composition with participants by role by respective disciplines.

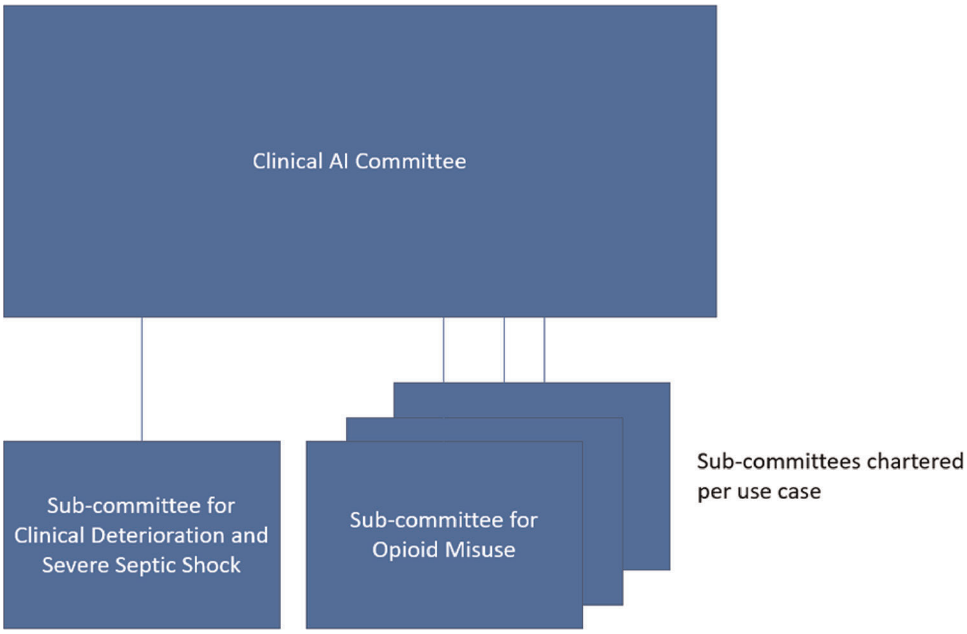

Clinical AI Committee
Sub-committee for Clinical Deterioration and Severe Septic Shock
Sub-committee for Opioid Misuse
Sub-committees chartered per use case

FIGURE 2
Clinical AI and predictive analytics committee and sub-committees.

3. Model output follows the five rights of Clinical Decision Support (CDS) * and is associated with interventions whenever possible.
4. Model monitoring (pilot or scale-out) includes statistical measures, operational metrics, relevant outcomes and reevaluation criteria, especially for calibration as absolute risk may change over time.
5. The basic principle of health care ethics autonomy, beneficence, justice and non-maleficence will be incorporated in all stages of model evaluation and validation. We aim to first do no harm with our AI-driven tools and ensure bioethical principles are integrated into our governance.

Predictive solution life cycle

A key aspect of appropriate governance is establishing a full life cycle for models. This includes processes for evaluation and potential adoption of models, monitoring to ensure they continuously meet the needs of all constituents, and appropriate processes for periodic reevaluation and decommissioning of models no longer needed or functioning correctly. Given the current institutional adoption of Lean methodology and specifically A3 thinking (5), we built our approval form starting with our institutional A3 project template, but added specific questions focused on relevant questions for AI implementation. A fuller description of the usefulness of Lean's FOCUS PDCA methodology for AI can be found in our previous work (6) with a toolkit available at www.hipxchange.org/ImplementPredictiveModels. **Supplementary Appendix 1** provides our model intake form, through which potential models are evaluated prior to approval. The intake form is designed to be completed in 2 stages. Basic model questions, in green, are designed to be filled out by the requestor prior to discussion with the Clinical AI and Predictive Analytics Committee. Once the committee has evaluated the use case, it can commission an algorithm subcommittee which provides the necessary expertise to complete the intake form in its entirety, which is necessary for model approval.

In addition, we developed a value stream beginning with the intake form through model re-evaluation over time. **Figure 3** depicts the life cycle of a typical predictive solution, from initial presentation to the Clinical AI and Predictive Analytics Committee to periodic review and update or decommissioning.

Results

Clear institutional apparatus for governance

At the time of this publication, the governance framework has overseen ten successful deployments, two successful retirements, and one successful non-deployment across nine applications. We expressly use the terminology of "deployment" and "retirement" as technical terms defined in the software and application development disciplines, where "deployment" refers to the promotion of the AI-driven application from a development environment into a production environment; and "retirement" is the removal of the application from a production environment after it is deemed to be no longer necessary. We distinguish this from the case of removing a solution from production due to errors or poor performance. The purpose of this technical terminology is to provide a necessary level of objectivity as it relates to endorsement, approvals, and IT change management. Applications include diverse uses of AI prediction for outputs including severe sepsis, clinical deterioration (7), physician panel weighting, COVID detection on radiographs (8), emergency department screening for falls prevention (9, 10), screening for opioid abuse (11), and Emergency Department crowding to drive adaptive staffing.

One key function of the governance framework is including all relevant AI applications. For AI applications which predated the current governance framework, there is an abbreviated process to grandfather these use cases into the current standard of oversight and transition *ad hoc* working groups to algorithm subcommittees which report up to the Committee. For AI-driven applications that are custom-built at UW Health, the University of Wisconsin, or involve a large-scale deployment, we have confidence that these are under the governance and oversight of this framework, due to the robust engagement and support of Informatics and IT within the current system. Another paradigm is the implementation of vendor-created models: these use cases are under governance particularly for vendor products that are explicitly marketed as an algorithm. However, we acknowledge that there may be use cases outside of the committee's awareness, especially for cases where the AI solution is embedded within a broader

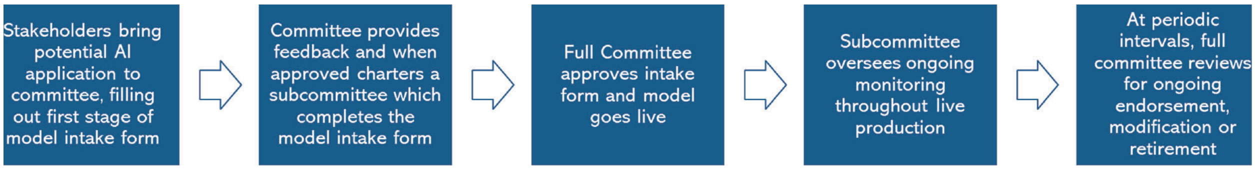

Stakeholders bring potential AI application to committee, filling out first stage of model intake form
Committee provides feedback and when approved charters a subcommittee which completes the model intake form
Full Committee approves intake form and model goes live
Subcommittee oversees ongoing monitoring throughout live production
At periodic intervals, full committee reviews for ongoing endorsement, modification or retirement

FIGURE 3
UW Health predictive model value stream.

product and is not marketed as an algorithm. Finally, we note that non-clinical use cases at our institution have adopted similar principles particularly the guiding principle of local validation of model performance.

Successful deployments spanning clinical and technology domains

We believe that one of the key drivers of our success has been comprehensive participation between clinical, operational, and IS stakeholders. While our IT professionals have a prominent role within our governance structure, AI application deployments are viewed as clinical projects analogous to other clinical initiatives in the hospital and our governance structure and use of A3 thinking mirrors that for purely clinical interventions such as clinical guideline development.

Process for ongoing monitoring to ensure performance

Once applications have been deployed, the algorithm sub-committees continue to meet on a pre-determined frequency that is compatible with its use case and to monitor the performance of the solution over subsequent years.

We maintain that ongoing monitoring has been successful as proven by two types of occurrences. The first type that has occurred is a successive deployment where the previous version of the AI application was replaced with a more performant solution that included cases spanning a different machine learning algorithm, an updated target or prediction outcome, or a re-trained model. This necessitates that the previous solution was actively monitored and that a new solution was also evaluated and validated with clear criteria regarding performance and acceptability.

The second type of occurrences are successful retirements, where an AI solution was removed from production after the solution delivered its intended value. We wish to clearly distinguish successful retirements from successful non-deployments. The latter indicates a situation where the AI solution was deemed to be non-performant prior to its use in clinical care and was never fully deployed in the production system. This differs from a successful retirement, where the AI solution was used in production as part of routine clinical care and performing appropriately. In these cases, the needs or requirements that the solution fulfilled have changed. For example, one of our successful applications was the use of a model to predict days with high emergency department volumes at one of our hospital sites, which was used to guide a decision to call in additional physician staff. While initially useful, this model slowly became less relevant as average daily volumes increased and daily staffing was increased, obviating the need for a call-in shift or the predictive model.

Mechanism to incorporate the equitable and ethical use of AI guidelines over time

To address the equitable and ethical use of AI, the membership of our institutional committee includes ethics expertise, including a prominent faculty member from the Law School, and staff from our office for Diversity, Equity, and Inclusion, and we maintain a line of communication with our medical bio-ethicists.

Drawing upon this expertise and membership, the guiding principles defined by the committee include guidelines for the equitable and ethical use of AI. We note that our guidelines also incorporate the evaluation of the intervention derived from the AI algorithm, which provides a more comprehensive determination of equitable and ethical impacts of how health interventions are administered across the system. **Supplementary Appendix 1**, the model intake form, shows explicit steps in the model vetting and monitoring process which are undertaken to ensure equity and evaluate for other potential ethical issues.

Our approach to the impactful enforcement of the equitable and ethical use of AI is to incorporate these guidelines as part of the same guiding principles that address technical and clinical guidelines. This is in contrast to treating the equity and ethics considerations as separate from other aspects of oversight. We believe this approach has been successful because it provides a clear charter for the sub-committees when applying the whole of the guiding principles to their use cases.

Interface with research

The committee specifically oversees AI applications which are instituted by the health system for the purpose of improving clinical care. We see this as a complementary role to research oversight; AI applications for research purposes are overseen by the IRB and research leadership. The committee is made aware of research-related IT build, and works with the IRB to ensure all AI applications are governed *via* either this clinical workflow or considered research.

Discussion

As adoption of AI applications in healthcare accelerates, there is an acute need for appropriate governance to address ethical, regulatory and trust concerns (12, 13). At the hospital level, effective governance offers the ability to specifically address these concerns while facilitating deployment and adoption (14). In our governance development, we have found that an effective system should not only be comprehensive, i.e., addressing all three domains as described in Challenges, but also adaptive, scaling appropriately with development of a program for implementing predictive solutions. This ensures that the level of oversight is proportional for each of the three domains at a given degree of program maturity. Our approach of adaptive governance evolved organically over time: we would likely have been unable to justify our current institutional committee without a number of extant solutions in need of oversight. At our current state we expect that our mechanisms will continue to evolve to meet organizational needs.

Faculty, in general have been supportive of the committee. While we expected some resistance to centralization of governance and a proscribed pathway for model deployment, these disadvantages seem to have been outweighed by the benefits of consistent expectations and process for all models. Research faculty have expressed favorable comments noting that introduces a consistent and supervised process for implementation of models after their development and validation.

While our current system developed organically, if we were to re-establish governance we would advocate a similar program of iterative building, allowing those involved in predictive model adoption to maintain flexibility early in the program and take advantage of gained institutional knowledge as it accrues. Key advantages of our current system of an oversight committee with federated small working groups include the breadth of recruited stakeholders and system scalability. Our current structure enables scaling by taking advantage of two tiers of governance. The institutional-level committee can and does charter multiple sub-committees as needed across multiple use cases to facilitate adoption and endorsement within each sub-committee's respective use cases. At the same time, the institution-level committee sets the guiding principles to enforce the consistency of standards and confidence of oversight while minimizing overhead. The goal is to meet the business needs of our health system while remaining cognizant of the AI guiding principles to prevent medical error and harm.

Data availability statement

This case study did not rely on specific data sets. Further inquiries can be directed to the corresponding authors.

Author contributions

FL and BWP conceived this manuscript. All authors contributed heavily to the development of the governance structure described. FL drafted the manuscript and all authors contributed substantially to its revision. All authors take responsibility for the content of this work.

Funding

This project was supported by UW Health. This work was also supported by AHRQ Grant Number R18HS027735 (Patterson, PI) and NIH/NIDA Grant Number R01DA051464 (Afshar, PI). The content is solely the responsibility of the authors and does not necessarily represent the official views of UW Health or the AHRQ.

Acknowledgments

The authors would like to thank UW Health and all the members and executive sponsors of the Clinical AI Committee.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2022.931439/full#supplementary-material.

References

1. Obermeyer Z, Emanuel EJ. Predicting the future — big data, machine learning, and clinical medicine. *N Engl J Med*. (2016) 375(13):1216–9. doi: 10.1056/NEJMp1606181

2. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med*. (2021) 385(3):283–6. doi: 10.1056/NEJMc2104626

3. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. *Nat Med*. (2019) 25(9):1337–40. doi: 10.1038/s41591-019-0548-6

4. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. (2019) 366(6464):447–53. doi: 10.1126/science.aax2342

5. Pyzdek T. Chapter 19: A3 thinking, mnaagement for professionals. In: *The lean healthcare handbook*. 2nd ed. Cham, Switzerland: Springer (2021). p. 223–43.

6. Smith MA, Adelaine S, Bednarz L, Patterson BW, Pothof J, Liao F. Predictive solutions in learning health systems: The critical need to systematize implementation of predictive models, actions, and interventions. *NEJM Catal Innov Care Deliv*. (2021) 2(5). doi: 10.1056/CAT.20.0650

7. Churpek MM, Yuen TC, Park SY, Meltzer DO, Hall JB, Edelson DP. Derivation of a cardiac arrest prediction model using ward vital signs*. *Crit Care Med*. (2012) 40(7):2102–8. doi: 10.1097/CCM.0b013e318250aa5a

8. Zhang R, Tie X, Qi Z, Bevins NB, Zhang C, Griner D, et al. Diagnosis of coronavirus disease 2019 pneumonia by using chest radiography: Value of artificial intelligence. *Radiology*. (2020) 2:e88–e97. doi: 10.1148/radiol.2020202944

9. Jacobsohn GC, Leaf M, Liao F, Maru AP, Engstrom CJ, Salwei ME, et al. Collaborative design and implementation of a clinical decision support system for automated fall-risk identification and referrals in emergency departments. *Healthcare*. (2022) 10(1):100598. doi: 10.1016/j.hjdsi.2021.100598

10. Patterson B, Engstrom C, Sah V, Smith MA, Mendonça EA, Pulia MS, et al. Training and interpreting machine learning algorithms to evaluate fall risk after emergency department visits. *Med Care*. (2019) 57(7):560–6. doi: 10.1097/MLR.0000000000001140

11. To D, Sharma B, Karnik N, Joyce C, Dligach D, Afshar M. Validation of an alcohol misuse classifier in hospitalized patients. *Alcohol*. (2020) 84:49–55. doi: 10.1016/j.alcohol.2019.09.008

12. Reddy S, Allan S, Coghlan S, Cooper P. A governance model for the application of AI in health care. *J Am Med Inform Assoc*. (2020) 27(3):491–7. doi: 10.1093/jamia/ocz192

13. World Health Organization. *Ethics and governance of artificial intelligence for health: WHO guidance*. Geneva, Switzerland: World Health Organization (2021). 150. https://www.who.int/publications/i/item/9789240029200

14. Marwaha JS, Landman AB, Brat GA, Dunn T, Gordon WJ. Deploying digital health tools within large, complex health systems: Key considerations for adoption and implementation. *NPJ digital medicine*. (2022) 5(1). doi: 10.1038/s41746-022-00557-1

frontiers | Frontiers in Digital Health

TYPE Perspective
PUBLISHED 02 September 2022
DOI 10.3389/fdgth.2022.958284

OPEN ACCESS

EDITED BY
Mark Sendak,
Duke Institute for Health Innovation,
United States

REVIEWED BY
Jonathan Lu,
Stanford University School of Medicine,
United States

*CORRESPONDENCE
Sharon E. Davis
sharon.e.davis.1@vumc.org

SPECIALTY SECTION
This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health

RECEIVED 31 May 2022
ACCEPTED 11 August 2022
PUBLISHED 02 September 2022

CITATION
Davis SE, Walsh CG and Matheny ME (2022)
Open questions and research gaps for monitoring and updating AI-enabled tools in clinical settings.
Front. Digit. Health 4:958284.
doi: 10.3389/fdgth.2022.958284

Open questions and research gaps for monitoring and updating AI-enabled tools in clinical settings

Sharon E. Davis[1]*, Colin G. Walsh[1,2,3] and Michael E. Matheny[1,2,4,5]

[1]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States, [2]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States, [3]Department of Psychiatry, Vanderbilt University Medical Center, Nashville, TN, United States, [4]Department of Biostatistics, Vanderbilt University Medical Center, Nashville, TN, United States, [5]Tennessee Valley Healthcare System VA Medical Center, Veterans Health Administration, Nashville, TN, United States

As the implementation of artificial intelligence (AI)-enabled tools is realized across diverse clinical environments, there is a growing understanding of the need for ongoing monitoring and updating of prediction models. Dataset shift—temporal changes in clinical practice, patient populations, and information systems—is now well-documented as a source of deteriorating model accuracy and a challenge to the sustainability of AI-enabled tools in clinical care. While best practices are well-established for training and validating new models, there has been limited work developing best practices for prospective validation and model maintenance. In this paper, we highlight the need for updating clinical prediction models and discuss open questions regarding this critical aspect of the AI modeling lifecycle in three focus areas: model maintenance policies, performance monitoring perspectives, and model updating strategies. With the increasing adoption of AI-enabled tools, the need for such best practices must be addressed and incorporated into new and existing implementations. This commentary aims to encourage conversation and motivate additional research across clinical and data science stakeholders.

KEYWORDS

dataset shift, model updating, machine learning, risk model surveillance, artificial intelligence

Introduction

As the implementation of artificial intelligence (AI)-enabled tools is realized across diverse clinical environments, there is a growing understanding of the need for ongoing monitoring and updating of prediction models (1–5). Beyond initial validation and local tailoring of models transported across settings, temporal deterioration in model accuracy after development has been documented across clinical domains and settings (6–10). Neither regression nor advanced machine learning algorithms are exempt from these temporal changes in performance (8, 11). Such performance drift degrades the clinical utility of AI-enabled tools, jeopardizes user trust, and poses safety concerns when insufficiently accurate predictions are used in decision-making (1, 6, 12, 13).

Dataset shift (14)—temporal changes in clinical practice, patient populations, and information systems—is well-documented as a source of performance drift and recognized as a challenge to the sustainability of AI-enabled tools in clinical care (6–8, 15–17). Model developers and system managers have access to a variety of approaches to address performance drift and underlying dataset shift in order to restore model performance to clinically acceptable levels. In some cases, model performance may be restored by correcting technical errors introduced by structural changes in information systems, such as implementation of revised data standards. However, in many cases where dataset shift is more nuanced and multifaceted, model updating through recalibration, retraining, or revision will be required. While best practices are well-established for training and validating new AI models (18), there is limited guidance on prospective validation and few best practices for model monitoring and updating.

In this paper, we highlight the need for maintaining clinical prediction models and discuss open questions regarding this critical aspect of the AI modeling lifecycle. First, we illustrate performance drift across models implemented in the production electronic health record (EHR) system at an academic medical center. Second, we discuss several open research questions and describe the nuances required for best practice guidance. Despite advances in continuous learning algorithms that evolve models as data accrue, such algorithms are subject to additional challenges and healthcare applications still predominantly rely on static models that will require periodic updating (19). Although we focus our discussion on updating static models, similar questions may arise around surveillance practices for continuous learning models.

Performance drift in operational models

Most studies documenting temporal model performance have been conducted in registry or research datasets rather than with operational data from models running in real-time clinical settings (7–9, 16). However, the transition from a retrospective research frame to real-time operational implementation may impact performance as input mappings change and the timing data availability shifts (20–22). To explore performance drift in an operational setting, we evaluated the performance of two models currently implemented in the production EHR system at Vanderbilt University Medical Center (VUMC): a non-proprietary, externally developed model predicting readmission (LACE+) (23) and a locally developed model predicting suicidal behaviors (Vanderbilt Suicide Attempt and Ideation Likelihood model, VSAIL) (24).

Table 1 provides an overview of each model, highlighting differences in modeling methods, training cohorts, and intended use. We extracted stored predictions calculated in real-time and outcomes associated with each prediction using data available in VUMC's EHR. For the LACE+ model, we note that this approach may undercount readmissions if patients were readmitted to a different medical facility. Monthly performance was evaluated using metrics relevant to each model's intended use. We measured the mean calibration of the LACE+ readmission model with the observed to expected outcome ratio (O:E) and clinical utility of the VSAIL suicidality model with the number needed to screen (NNS; the inverse of positive predictive value).

The LACE+ model, locally calibrated to the VUMC population, sustained performance over the evaluation period (**Figure 1A**). Monitoring highlighted the importance of distinguishing noise from both informative local change in performance and true model deterioration. Over the first 2.5 years, variability in observed O:Es did not follow a significant trend. In the last year of evaluation, however, there may be a trend toward lower O:Es. Depending on the use case, this declining O:E could be seen as indicating improved local quality (i.e., reducing readmissions) or increasing miscalibration. We note that O:E, a crude measure of calibration, may conceal calibration drift within clinically important risk ranges (25). VSAIL maintained a relatively stable NNS during the first year of implementation (median monthly NNS = 19), with the NNS abruptly increasing in February 2021 (median monthly NNS = 136);

TABLE 1 Prediction models evaluated for temporal validation of real-time scores generated within a production electronic medical record system.

Details	LACE+	VSAIL
Outcome	30-day readmission	30-day suicidal ideation or attempt
Intended use	Quality benchmarking using predicted risk of readmission calculated at discharge	Clinical decision support delivered at arrival for inpatient and outpatient encounters
Development setting	Patients from multiple hospitals in Ontario, Canada	VUMC patient population
Modeling approach	Logistic regression	Random forest
Evaluation period	January 2018 through March 2022	December 2019 through January 2022

VUMC, Vanderbilt University Medical Center.

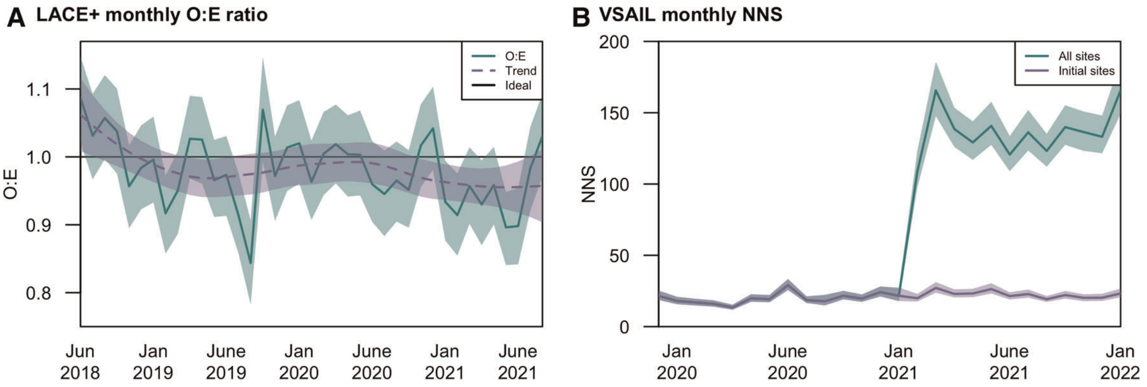


FIGURE 1
Temporal performance at Vanderbilt University Medical Center of the (A) LACE+ readmission model in terms of mean calibration (O:E); and (B) VSAIL suicidality model in terms of number needed to screen (NNS).

Figure 1B). This shift corresponds to operational changes in implementation, with the model being applied to a much broader patient population. Within the original population, VSAIL's NNS remained stable (median monthly NNS = 22). The higher NNS in the broader population may still be feasible but should be considered in the implementation team's cost-benefit analysis and may warrant further investigation of performance in select clinical settings or subpopulations. These findings illustrate performance drift in a single health system's EHR and contribute to the mounting evidence that AI-enabled tools require long-term strategies to understand performance trajectories and maintain utility.

Research and best practice gaps in model maintenance

Despite concerns over the long-term stability of model performance, health systems lack generalizable guidance for operationalizing post-implementation maintenance strategies. To develop guidance and establish best practices, additional research and debate are needed in three focus areas: model maintenance policies, performance monitoring perspectives, and model updating strategies (Table 2).

Maintenance policies

Oversight policies at the health system level could facilitate the maintenance of a portfolio of models by defining a consistent, systematic groundwork for sustaining both new and existing AI-enabled tools. System-level policies can also inform use case parameters to consider when establishing model-specific maintenance plans.

How should model ownership impact local control over maintenance?

Health systems certainly have a right and duty to monitor the local performance of the AI models they implement, regardless of where those models originated. However, how to address deteriorations in performance is complicated by model ownership and licensing restrictions. For models developed in-house and local implementations of models in the public domain, health systems have full control over maintenance approaches and may consider the full spectrum of updating methods. At VUMC, the local VSAIL model will be retrained using more recent data and subsequently maintained through a tailored data-driven surveillance approach. When models are developed in collaboration across health systems, best practices could guide collaborative updating and establishment of parameters for local model adjustments.

Updating proprietary models is particularly challenging, despite locally documented drift having required the deactivation of proprietary AI-enabled clinical tools (6). Licenses may restrict updating by not permitting local model recalibration or retraining. Updating options may be further limited by inadequate documentation of training methods (26). Proactive updating of proprietary models by model owners, such as semi-annual updates of the National Surgical Quality Improvement Program (ACS NSQIP) risk models, may alleviate some, but not all, of the need for local updating options. Health systems, national organizations, and policymakers should advocate for more complete documentation of proprietary models and increased access to updating options. This may include the relaxation of local updating restrictions; clear documentation of owner-driven maintenance plans; and proactive, transparent dissemination of updated models to all customers. Enabling such expectations of model owners will require more detailed and

TABLE 2 Overview of gaps in best practices for model maintenance.

Domain	Gaps/Needs
Maintenance policies	
How should model ownership impact local control over maintenance?	• Policies establishing updating expectations of proprietary models • Clarity and fairness of local updating opportunities of proprietary models • Prototypes for establishing collaborative updating of multi-system owned models
How do we ensure comparable performance across demographic groups is sustained during the maintenance phase?	• Guidance on whether and when changes in model fairness warrant pausing AI-enabled tools • Methods for addressing performance fairness drift when model performance deteriorates differentially across subpopulations
How do we communicate model changes to end users and promote acceptance?	• Design of effective communication strategies for warning end users of model performance drift and informing users when updated models are implemented • Guidance on aligning messaging with end-user AI literacy
Performance monitoring	
At what level should model performance be monitored and maintained?	• Guidance on aligning monitoring and maintenance with use case needs • Recommendations for handling monitoring in smaller health systems, including determining minimum sample size and methods for collaborative monitoring • Policies supporting collaborative model maintenance in low data resource settings • Guidance on managing interim periods of local performance drift between releases of proprietary models that cannot be locally updated
What aspects of performance should be monitored?	• Generalization recommendations on frequency and sample sizes for measuring performance across a variety of metrics • Customizable and expandable tools to monitor a matrix of metrics • Guidelines for aligning metrics of interest with use case needs
How do we define meaningful changes in performance?	• Framework for selecting drift detection methods • Guidance on establishing clinically acceptable ranges of performance and defining clinically relevant decision boundaries • Methods for tailoring drift detection algorithms to detect a clinically important change
Are there other aspects of AI models that we should monitor, in addition to performance?	• Approaches to systematically surveil external features that may impact model inputs and for monitoring input data distributions • Guidance on when to update in response to changes in model inputs if performance remains stable • Systems for disseminating information on changes anticipated to affect common AI models
Model updating	
What updating approaches should be considered?	• Approaches to optimizing update method selection based on performance characteristics most relevant to use case needs • Expanded suite of testing procedures options for more updating methods and increased computational efficiency • Guidance on defining acceptable performance and methods to determine which updating methods, if any, restore acceptable performance
Should clinically meaningful or statistically significant changes in performance guide updating practice?	• Guidance on whether to update models when statistically significant improvement is possible but updating would not provide a clinically meaningful improvement • Methods for comparing updating options that incorporate tests for both statistical and clinical significance • Recommendations for decision-making in cases where available updating methods do not restore performance to acceptable levels
How do we handle biased outcome feedback after model implementation?	• Recommendations for assessing feedback from effective AI-enabled interventions • Methods for model development, validation, and updating that are robust to confounding by intervention

consistent guidance on model updating practices covering the concerns described throughout this paper.

How do we ensure comparable performance across demographic groups is sustained during the maintenance phase?

Model fairness is now recognized as a critical element of clinical AI models (27, 28). While model fairness comprises a broad set of concerns regarding implementation practices, user uptake and application, and sociotechnical contexts of use (29), fairness also requires models to perform similarly across demographic groups. Establishing initial comparable model performance across subpopulations and subsequently maintaining comparable performance within these groups is thus critical to ensuring model fairness. Novel metrics for evaluating algorithmic fairness across subpopulations are providing insight during model validation and selection of models for implementation in clinical tools (30, 31). A clear next step is to incorporate these new metrics and performance within subpopulations into model monitoring to evaluate fairness over time. This poses new questions regarding how to handle the potential for fairness drift, defined as differential performance drift across subpopulations. Researchers and policymakers will need to address tests for temporal changes in fairness metrics; methods for updating models experiencing fairness drift that prioritize equitable utility for all patients; and whether and when changes in model fairness warrant pausing AI-enabled tools to avoid creating or exacerbating disparities.

How do we communicate model changes to end users and promote acceptance?

Open communication between modeling teams and clinical end users is essential to the monitoring and maintenance phase of the AI lifecycle. End users may identify failing AI-enabled tools before performance monitors detect changes in accuracy. They may also provide insight when models are no longer useful from a clinical perspective even with sustained performance, allowing tools to be de-implemented or revised as needed. At the same time, modeling teams should establish policies for disseminating information about model updates to end-users, whether updating is driven by end-user concerns, local model maintenance efforts, or new releases of proprietary models. Such communication, while particularly important for reestablishing trust in models updated in response to end-user concerns, is relevant for all updates. Model maintenance programs need to include specific strategies for this bidirectional communication. Such engagement and transparency regarding model maintenance may also increase acceptance of AI more broadly by assuring users that models are actively being curated, monitored, and assessed with an eye to promoting utility and safety.

The appropriate mode of communication and level of detail provided about model updates are likely to use case-dependent. The ACS NSQIP surgical risk calculator, for example, displays a banner message highlighting recent updates, setting expectations for any noticeable changes in predictions, and eliciting feedback if concerns arise (32). Extensive model revisions or reimplementation of a paused model with restored performance may require more explanation than a banner message can effectively convey. Workflow and communication experts will be key collaborators in designing best practices for disseminating information on model updates. These best practices will likely need to evolve as the health care workforce becomes better trained in AI.

Performance monitoring perspectives

Ongoing monitoring provides necessary insight into model stability and can alert model managers to concerning performance trends in need of intervention (3, 33, 34). However, insights from monitoring require careful determinations of how model performance is defined and evaluated.

At what level should model performance be maintained?

AI models, even when operationalized to meet the needs of a specific health system, may need to be monitored and updated locally, regionally, or nationally. Key features to consider in determining the appropriate level of model maintenance include use case goals, model ownership, and data and analysis resources.

Our understanding of best practices is well-defined in terms of use case and the level of model maintenance. For benchmarking models in quality evaluations, maintenance should be centralized at the largest relevant scope. Stabilizing the performance of quality-oriented models at higher levels imbues local performance deviations with information about variations in care and allows facilities to validly interpret performance trends as indicating improvement or deterioration of local performance over time. For AI-enabled tools aimed at clinical decision-making and population management, individual predictions should be well-calibrated to ensure utility and benefit to patients (13). As a result, more localized monitoring and maintenance are appropriate.

Unfortunately, practical considerations may require centralized monitoring and updating at regional or national scales even when local performance would typically be prioritized. Ownership and licensing requirements of proprietary models may preclude updating models to optimize local performance. Guidance on how to assess and handle local drift in light of such restrictions is necessary to trigger pauses in model implementations when local monitoring efforts reveal concerning performance drift; facilitate communication with end users about paused models and support end users' information needs during such pauses, and promote timely reporting of issues to model owners.

When local updating is permissible, monitoring and updating remain a challenge for small organizations where data volumes and analytic resources may be limited. Insufficient sample sizes can lead to highly variable performance during monitoring and limit the ability to distinguish performance drift from noise. Smaller organizations, as well as their larger peers, should leverage recent studies by Riley et al. to assess whether sufficient sample sizes are available to validate binary (35), time-to-event (36), and continuous models (35). Recalibration, retraining, and model revision also require sufficient sample sizes (37) and dedicated data science teams that may not be feasible for all organizations. One solution would be to explore whether health information exchanges could be leveraged for collaborative monitoring and updating where local resources are insufficient. Broader research and policy discussions are needed as we think creatively about such multi-level, coordinated efforts to ensure the benefits of predictive tools are available and practical for health care organizations serving all communities.

What aspects of performance should be monitored?

While some metrics appear more robust to dataset shift, performance drift has been documented in measures of discrimination, calibration, and clinical utility (7, 8, 10, 16, 38). Monitoring metrics relevant to an AI-enabled tool's use case is critical to understanding whether changes in performance warrant updating or whether updating may have little impact on model use and outcomes. For example, the number needed to screen was identified by the VSAIL team as the target metric for monitoring and stabilizing model performance as this impacts the cost-benefit analysis of clinics adopting the tool (39). For models deployed in diverse clinical contexts or across multiple tools, tracking a matrix of performance measures would provide insights supporting a variety of user perspectives (12, 40). Monitoring recommendations should thus include components that are agnostic to the performance metrics under consideration (e.g., selection of measurement), as well as components regarding metric selection.

How do we define meaningful changes in performance?

Monitoring performance alone is insufficient; model managers need to be able to determine when observed deterioration in performance warrants intervention. Drift detection methods surveil temporal performance to alert users to statistically significant changes (41, 42) and have been applied to monitoring clinical prediction models. (34, 38) Methods vary in their ability to handle multiple forms and speeds of performance drift, as well as in their applicability to clinical contexts where calibration is of interest (43). Best practice recommendations will need to provide a decision framework for selecting between drift detection approaches, including considerations of whether detection algorithms are model-independent; can handle data streams of individual or batched observations, and are flexible in their ability to monitor prediction errors using a variety of metrics.

We note small differences in performance may be detected by the statistical tests underlying drift detection algorithms. However, statistically significant differences in performance may not directly translate into clinically meaningful differences. In such cases, users may question the value of updating or pausing a model in response to detections of small statistically significant, but not clinically important performance drift. The magnitude of acceptable inaccuracy and performance variability likely varies by use case. For example, performance drift is most likely to impact clinical utility when the calibration of predictions near clinically relevant decision thresholds or near classification cut-points deteriorates. Understanding whether, when, and how performance drift affects the clinical utility of predictions for decision-making is key to detecting meaningful changes in monitored models. Defining and measuring clinically acceptable performance and defining clinically relevant decision boundaries remains an open area of research. Subsequent research and guidance will need to address tailoring drift detection algorithms to place more import on clinically important changes in model performance.

Are there other aspects of AI models that we should monitor?

In addition to performance metrics, the inputs of AI models could be monitored. This may involve evaluating data streams for changes in predictor distributions and associations (17), as well as establishing teams to actively evaluate external influences in clinical guidelines, software systems, data standards, and health care policies (6). Tracking external influences would allow teams to recognize structural changes that could render a model unreliable and plan customized updating approaches. Changes in data stream features, however, may not necessitate updating unless and until they affect the model accuracy in clinically meaningful ways. Best practices will need to address integrating insights from performance monitoring and evaluations of factors impacting model inputs to promote stable performance while efficiently and conservatively updating models. Additional research could investigate strategies for monitoring these non-performance aspects of AI models and policies for disseminating information across health systems when new practices are anticipated to impact widely adopted models.

Updating strategies

When updating is initiated by pre-established schedules or detected performance drift, model managers must choose between a range of updating methods – from recalibration to retraining to model revision. As not all methods will be feasible, permissible, or successful in all situations, research and recommendations are needed to guide updating practice.

What updating approaches should be considered?

Although retraining with a cohort of recent observations may be established practice, this approach fails to build on the knowledge encoded in existing models, can be susceptible to overfitting, and may not improve performance above that achieved through recalibration (11, 44–47). For health systems with smaller populations, concerns regarding performance instability when retraining complex models may be more pronounced. Several methods have been developed to compare updating approaches on a particular cohort and recommend the approach that most improves accuracy (17, 45, 46). These methods, however, test for statistically rather than clinically significant differences across potential updates and do not consider whether the recommended update sufficiently restores performance. As methods for establishing clinically relevant decision thresholds mature, testing procedures for selecting updating methods could be implemented with weighted scoring rules to emphasize accuracy in critical regions. Future research should consider expanding options for optimizing decisions using varied performance metrics; increasing test efficiency, particularly for computationally intensive models; methods for evaluating whether updating provides clinically meaningful improvement; and recommendations for cases in which available updating methods do not restore models to acceptable levels of accuracy.

How do we handle biased outcome feedback after implementation?

Model updating with current recalibration, retraining, and model revision methods has been developed, evaluated, and applied primarily in research databases. In production systems, interactions between users and AI-enabled decision support tools will, if successful, alter treatment decisions and improve patient outcomes. As a result, the observed data in production systems will be biased and updates using these biased data may reduce future model utility by updating away useful signals (48). These feedback loops created by successful clinical AI tools pose new challenges to updating practice that requires additional methodological research to better characterize the problem; to distinguish between dataset shift and performance changes due to model interventions; and to develop novel algorithms and updating approaches that are robust to confounding by intervention.

Conclusion

The clinical AI lifecycle is incomplete without components to monitor and stabilize accuracy in evolving clinical environments. Despite the diverse landscape of AI-enabled tools, common challenges to model maintenance impact new and existing implementations regardless of clinical domain and underlying modeling algorithms. Methods development for model monitoring and updating is accelerating, yet open questions for the design of maintenance programs, those described here and more, require additional research and scientific consensus to devise best practices. Establishing best practices is critical to designing AI-enabled tools that deliver reliable predictions, promote adoption, and realize the promise of AI to improve patient care.

Data availability statement

The datasets presented in this article are not readily available because VUMC patient data used in this study are not publicly available. Requests to access the datasets should be directed to Sharon Davis, sharon.e.davis.1@vumc.org.

Ethics statement

The studies involving human participants were reviewed and approved by Vanderbilt University Medical Center's Institutional Review Board. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

SED and MEM developed the initial themes of this manuscript. SED and CGW conducted data management and analysis. SED drafted the manuscript. All authors contributed to the article and approved the submitted version.

Funding

CW and the Vanderbilt Suicide Attempt and Ideation Likelihood model (VSAIL) were supported by funding from the National Institutes of Health (R01 MH121455; R01 MH120122; R01 MH116269), the Military Suicide Research Consortium (W81XWH-10-2-0181), and Vanderbilt University Medical Center's Evelyn Selby Stead Fund for Innovation.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Amarasingham R, Patzer RE, Huesch M, Nguyen NQ, Xie B. Implementing electronic health care predictive analytics: considerations and challenges. *Health Aff.* (2014) 33(7):1148–54. doi: 10.1377/hlthaff.2014.0352

2. Smith J. Setting the agenda: an informatics-led policy framework for adaptive CDS. *J Am Med Inform Assoc.* (2020) 27(12):1831–3. doi: 10.1093/jamia/ocaa239

3. Matheny ME, Thadaney Israni S, Ahmed M, Whicher D. *Artificial intelligence in health care: the hope, the hype, the promise, the peril.* Washington, DC: National Academy of Medicine (2019).

4. Jenkins DA, Martin GP, Sperrin M, Riley RD, Debray TPA, Collins GS, et al. Continual updating and monitoring of clinical prediction models: time for dynamic prediction systems? *Diagn Progn Res.* (2021) 5(1):1. doi: 10.1186/s41512-020-00090-3

5. Petersen C, Smith J, Freimuth RR, Goodman KW, Jackson GP, Kannry J, et al. Recommendations for the safe, effective use of adaptive CDS in the US healthcare system: an AMIA position paper. *J Am Med Inform Assoc.* (2021) 28(4):677–84. doi: 10.1093/jamia/ocaa319

6. Finlayson SG, Subbaswamy A, Singh K, Bowers J, Kupke A, Zittrain J, et al. The clinician and dataset shift in artificial intelligence. *N Engl J Med.* (2021) 385(3):283–6. doi: 10.1056/NEJMc2104626

7. Hickey GL, Grant SW, Murphy GJ, Bhabra M, Pagano D, McAllister K, et al. Dynamic trends in cardiac surgery: why the logistic EuroSCORE is no longer suitable for contemporary cardiac surgery and implications for future risk models. *Eur J Cardiothorac Surg.* (2013) 43(6):1146–52. doi: 10.1093/ejcts/ezs584

8. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc.* (2017) 24(6):1052–61. doi: 10.1093/jamia/ocx030

9. Minne L, Eslami S, De Keizer N, De Jonge E, De Rooij SE, Abu-Hanna A. Effect of changes over time in the performance of a customized SAPS-II model on the quality of care assessment. *Intensive Care Med.* (2012) 38(1):40–6. doi: 10.1007/s00134-011-2390-2

10. Wong A, Cao J, Lyons PG, Dutta S, Major VJ, Otles E, et al. Quantification of sepsis model alerts in 24 US hospitals before and during the COVID-19 pandemic. *JAMA Netw Open.* (2021) 4(11):e2135286. doi: 10.1001/jamanetworkopen.2021.35286

11. Davis SE, Greevy RA, Lasko TA, Walsh CG, Matheny ME. *Comparison of prediction model performance updating protocols: using a data-driven testing procedure to guide updating. Proceedings of the AMIA Annual Symposium* Bethesda, MD: American Medical Informatics Association (2019). pp. 1002–10.

12. Jiang X, Osl M, Kim J, Ohno-Machado L. Calibrating predictive model estimates to support personalized medicine. *J Am Med Inform Assoc.* (2012) 19(2):263–74. doi: 10.1136/amiajnl-2011-000291

13. Van Calster B, Vickers AJ. Calibration of risk prediction models: impact on decision-analytic performance. *Med Decis Making.* (2015) 35(2):162–9. doi: 10.1177/0272989X14547233

14. Quinonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence N. *Dataset shift in machine learning.* Cambridge, MA: The MIT Press (2009).

15. Luijken K, Wynants L, van Smeden M, Van Calster B, Steyerberg EW, Groenwold RHH, et al. Changing predictor measurement procedures affected the performance of prediction models in clinical examples. *J Clin Epidemiol.* (2020) 119:7–18. doi: 10.1016/j.jclinepi.2019.11.001

16. Davis SE, Lasko TA, Chen G, Matheny ME. *Calibration drift among regression and machine learning models for hospital mortality. Proceedings of the AMIA Annual Symposium* Bethesda, MD: American Medical Informatics Association (2017). pp. 625–34.

17. Guo LL, Pfohl SR, Fries J, Posada J, Fleming SL, Aftandilian C, et al. Systematic review of approaches to preserve machine learning performance in the presence of temporal dataset shift in clinical medicine. *Appl Clin Inform.* (2021) 12(4):808–15. doi: 10.1055/s-0041-1735184

18. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res.* (2016) 18(12):e323. doi: 10.2196/jmir.5870

19. Jenkins DA, Sperrin M, Martin GP, Peek N. Dynamic models to predict health outcomes: current status and methodological challenges. *Diagn Prognostic Res.* (2018) 2:23.

20. Morse KE, Brown C, Fleming S, Todd I, Powell A, Russell A, et al. Monitoring approaches for a pediatric chronic kidney disease machine learning model. *Appl Clin Inform.* (2022) 13(2):431–8. doi: 10.1055/s-0042-1746168

21. Walsh CG, Johnson KB, Ripperger M, Sperry S, Harris J, Clark N, et al. Prospective validation of an electronic health record-based, real-time suicide risk model. *JAMA Netw Open.* (2021) 4(3):e211428. doi: 10.1001/jamanetworkopen.2021.1428

22. Otles E, Oh J, Li B, Bochinski M, Joo H, Ortwine J, et al. *Mind the performance gap: examining dataset shift during prospective validation. Proceedings of the 6th Machine Learning for Healthcare Conference* Proceedings of Machine Learning Research (2021). pp. 506–34.

23. van Walraven C, Wong J, Forster AJ. LACE+ index: extension of a validated index to predict early death or urgent readmission after hospital discharge using administrative data. *Open Med.* (2012) 6(3):e80–90.

24. Walsh C, Ribeiro J, Franklin J. Predicting risk of suicide attempts over time through machine learning. *Clin Psychol Sci.* (2017) 5(3):457–69. doi: 10.1177/2167702617691560

25. Van Calster B, Nieboer D, Vergouwe Y, De Cock B, Pencina MJ, Steyerberg EW. A calibration hierarchy for risk models was defined: from utopia to empirical data. *J Clin Epidemiol.* (2016) 74:167–76. doi: 10.1016/j.jclinepi.2015.12.005

26. Lu JH, Callahan A, Patel BS, Morse KE, Dash D, Pfeffer MA, et al. Assessment of adherence to reporting guidelines by commonly used clinical prediction models from a single vendor: a systematic review. *JAMA Netw Open.* (2022) 5(8):e2227779. doi: 10.1001/jamanetworkopen.2022.27779

27. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med.* (2018) 178(11):1544–7. doi: 10.1001/jamainternmed.2018.3763

28. Paulus JK, Kent DM. Predictably unequal: understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *NPJ Digit Med.* (2020) 3:99. doi: 10.1038/s41746-020-0304-9

29. Selbst A, Boyd D, Friedler S, Venkatasubramanian S, Vertesi J., *Fairness and abstraction in sociotechnical systems. ACM Conference of Fairness, Accountability, and Transparency* New York, NY: Association for Computing Machinery (2019). pp. 59–68.

30. Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. *J Biomed Inform.* (2021) 113:103621. doi: 10.1016/j.jbi.2020.103621

31. Beutel A, Chen J, Doshi T, Qian H, Woodruff A, Luu C, et al. *Putting fairness principles into practice: challenges, metrics, and improvements. 2019 AAAI/ACM Conference on AI, Ethics, and Society* New York, NY: Association for Computing Machinery (2019). pp. 453–9.

32. ACS NSQIP Surgical Risk Calculator. Available at: https://riskcalculator.facs.org/RiskCalculator/index.jsp. (accessed April 30, 2022).

33. Jung K, Kashyap S, Avati A, Harman S, Shaw H, Li R, et al. A framework for making predictive models useful in practice. *J Am Med Inform Assoc.* (2021) 26(6) 1149–58.

34. Davis SE, Greevy Jr RA., Lasko TA, Walsh CG, Matheny ME. Detection of calibration drift in clinical prediction models to inform model updating. *J Biomed Inform.* (2020) 112:103611. doi: 10.1016/j.jbi.2020.103611

35. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, van Smeden M, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med.* (2021) 40(19):4230–51. doi: 10.1002/sim.9025

36. Riley RD, Collins GS, Ensor J, Archer L, Booth S, Mozumder SI, et al. Minimum sample size calculations for external validation of a clinical prediction model with a time-to-event outcome. *Stat Med*. (2022) 41 (7):1280–95. doi: 10.1002/sim.9275

37. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med*. (2004) 23(16):2567–86. doi: 10.1002/sim.1844

38. Minne L, Eslami S, de Keizer N, de Jonge E, de Rooij SE, Abu-Hanna A. Statistical process control for monitoring standardized mortality ratios of a classification tree model. *Methods Inf Med*. (2012) 51(4):353–8. doi: 10.3414/ME11-02-0044

39. Ross EL, Zuromski KL, Reis BY, Nock MK, Kessler RC, Smoller JW. Accuracy requirements for cost-effective suicide risk prediction among primary care patients in the US. *JAMA Psychiatry*. (2021) 78(6):642–50. doi: 10.1001/jamapsychiatry.2021.0089

40. Steyerberg EW, Vickers AJ, Cook NR, Gerds T, Gonen M, Obuchowski N, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology*. (2010) 21(1):128–38. doi: 10.1097/EDE.0b013e3181c30fb2

41. Gama J, Žliobaitė I, Bifet A, Pechenizkiy M, Bouchachia A. A survey on concept drift adaptation. *ACM Comput Surv (CSUR)*. (2014) 46(4):44. doi: 10.1145/2523813

42. Bifet A, Gavalda R. *Learning from time-changing data with adaptive windowing*. *Proceedings of the 2007 SIAM International Conference on Data Mining*. Philadephia, PA: Society for Industrial and Applied Mathematics (2007). pp. 443–8.

43. Benneyan JC, Lloyd RC, Plsek PE. Statistical process control as a tool for research and healthcare improvement. *BMJ Qual Saf*. (2003) 12(6):458–64. doi: 10.1136/qhc.12.6.458

44. Toll DB, Janssen KJ, Vergouwe Y, Moons KG. Validation, updating and impact of clinical prediction rules: a review. *J Clin Epidemiol*. (2008) 61 (11):1085–94. doi: 10.1016/j.jclinepi.2008.04.008

45. Vergouwe Y, Nieboer D, Oostenbrink R, Debray TP, Murray GD, Kattan MW, et al. A closed testing procedure to select an appropriate method for updating prediction models. *Stat Med*. (2017) 36(28):4529–39. doi: 10.1002/sim.7179

46. Davis SE, Greevy RA, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. *J Am Med Inform Assoc*. (2019) 26(12):1448–57. doi: 10.1093/jamia/ocz127

47. Su TL, Jaki T, Hickey GL, Buchan I, Sperrin M. A review of statistical updating methods for clinical prediction models. *Stat Methods Med Res*. (2018) 27(1):185–97. doi: 10.1177/0962280215626466

48. Lenert MC, Matheny ME, Walsh CG. Prognostic models will be victims of their own success, unless. *J Am Med Inform Assoc*. (2019) 26(12):1645–50.

frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 05 September 2022
DOI 10.3389/fdgth.2022.932123

From compute to care: Lessons learned from deploying an early warning system into clinical practice

Chloé Pou-Prom[1]*, Joshua Murray[2], Sebnem Kuzulugil[1], Muhammad Mamdani[1,3,4,5,6,7†] and Amol A. Verma[1,3,4†]

[1]Data Science and Advanced Analytics, St. Michael's Hospital, Unity Health Toronto, Toronto, ON, Canada, [2]Department of Statistics, University of Toronto, Toronto, ON, Canada, [3]Department of Medicine, Faculty of Medicine, University of Toronto, Toronto, ON, Canada, [4]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, ON, Canada, [5]Dalla Lana School of Public Health, University of Toronto, Toronto, ON, Canada, [6]Vector Institute, Toronto, ON, Canada, [7]Leslie Dan Faculty of Pharmacy, University of Toronto, Toronto, ON, Canada

Background: Deploying safe and effective machine learning models is essential to realize the promise of artificial intelligence for improved healthcare. Yet, there remains a large gap between the number of high-performing ML models trained on healthcare data and the actual *deployment* of these models. Here, we describe the deployment of CHARTwatch, an artificial intelligence-based early warning system designed to predict patient risk of clinical deterioration.

Methods: We describe the end-to-end infrastructure that was developed to deploy CHARTwatch and outline the process from data extraction to communicating patient risk scores in real-time to physicians and nurses. We then describe the various challenges that were faced in deployment, including technical issues (e.g., unstable database connections), process-related challenges (e.g., changes in how a critical lab is measured), and challenges related to deploying a clinical system in the middle of a pandemic. We report various measures to quantify the success of the deployment: model performance, adherence to workflows, and infrastructure uptime/downtime. Ultimately, success is driven by end-user adoption and impact on relevant clinical outcomes. We assess our deployment process by evaluating how closely we followed existing guidance for good machine learning practice (GMLP) and identify gaps that are not addressed in this guidance.

Results: The model demonstrated strong and consistent performance in real-time in the first 19 months after deployment (AUC 0.76) as in the silent deployment heldout test data (AUC 0.79). The infrastructure remained online for >99% of time in the first year of deployment. Our deployment adhered to all 10 aspects of GMLP guiding principles. Several steps were crucial for deployment but are not mentioned or are missing details in the GMLP principles, including the need for a silent testing period, the creation of robust downtime protocols, and the importance of end-user engagement. Evaluation for impacts on clinical outcomes and adherence to clinical protocols is underway.

Conclusion: We deployed an artificial intelligence-based early warning system to predict clinical deterioration in hospital. Careful attention to data

Check for updates

OPEN ACCESS

EDITED BY
David Vidal,
Mayo Clinic, United States

REVIEWED BY
Xia Jing,
Clemson University, United States
Danny Eytan,
Technion Israel Institute of Technology, Israel

*CORRESPONDENCE
Chloé Pou-Prom
Chloe.Pou-Prom@unityhealth.to

†These authors share senior authorship.

SPECIALTY SECTION
This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health

RECEIVED 29 April 2022
ACCEPTED 08 August 2022
PUBLISHED 05 September 2022

CITATION
Pou-Prom C, Murray J, Kuzulugil S, Mamdani M and Verma AA (2022) From compute to care: Lessons learned from deploying an early warning system into clinical practice.
Front. Digit. Health 4:932123.
doi: 10.3389/fdgth.2022.932123

infrastructure, identifying problems in a silent testing period, close monitoring during deployment, and strong engagement with end-users were critical for successful deployment.

KEYWORDS

machine learning, healthcare, deployment, early warning system, clinical pathway

Introduction

Despite advancements in machine learning algorithms for solving healthcare problems, there still remains a gap between the number of developed algorithms and the number of successful deployments (1, 2).

Problems can arise at any stage of deployment (3). Prior to model development, unclear problem definition is often cited as a barrier to successful deployment (4). Then, during model development, the training data can be biased, either due to missingness of vulnerable populations, small sample size, or erroneous data (5). When transitioning to production data, there can be a drop in model performance from test data to production data, which may result from changes in data formats, timing, and context (3). Problems can also arise with out-of-distribution generalization and incorrect feature attribution (for example, if clinical protocols or target populations change over time) (6). If a model makes it to the deployment phase, end-user engagement is a crucial facilitator of, or barrier to, successful uptake. Introducing a clinical team to a ML model may require changes in workflow and change management. Ensuring that end-users correctly use a deployed product is difficult if there is no buy-in or trust (7).

Because of the issues listed above, there are few successful deployments of ML in healthcare settings. The scarcity of deployments means there are no widely accepted "best practices" or standards by which to evaluate the success of a deployment. Recent work has looked at assessing the quality of the data that goes into model through the creation of "Datasheets for Datasets" (8). In the past year, Health Canada, the U.S. Food and Drug Administration (FDA), and the United Kingdom's Medicines and Healthcare products Regulatory Agency (MHRA) have released the Good Machine Learning Practice (GMLP) guiding principles, a document providing 10 principles to address deployment of healthcare algorithms (9). More recently, the DECIDE-AI steering committee have released DECIDE-AI, a set of guidelines and checklists meant for early live clinical evaluation.

TABLE 1 Example of measured labs and vitals in the CHARTwatch training dataset. For these vitals and labs, we report their mean values, the 1st quantile (Q01), the 99th quantile (Q99), minimum value (Min) and maximum value (Max). The minimum and maximum values often fall outside of the range of biologically possible values (e.g., a maximum body temperature value of 6932 °C).

Feature	Mean	Q01	Q99	Min	Max
Vital—temperature	36.91	34.8	38.3	0	6932
Vital—diastolic blood pressure	71.49	47	101	0	173
Vital—systolic blood pressure	129.89	85	183	1	16,070
Vital—respirations	20.53	15	28	0	20,147
Lab—troponin	0.28	−1.9	5.58	−1.9	7.13
Lab—HBA1	−2.73	−3.14	−1.9	−3.3	2.29
Lab—glucose random	1.95	1.22	3.29	0	4.51
Lab—Hemoglobin	106.09	63	160	1	214
Lab—Basophils	0.03	0	0.13	0	2.69
Lab—Alanine Aminotransferease	3.38	1.61	7.08	1.61	8.78

In Fall 2020, we deployed CHARTwatch to the General Internal Medicine (GIM) ward at St. Michael's Hospital, an inner city teaching hospital in Canada (2, 10). The GIM ward currently holds 78 beds and receives approximately 4,000 admissions each year. Here, we describe in detail, the system's infrastructure and assess the success of our deployment through quantitative metrics (such as model performance, end-user engagement, and adherence to workflows) and by comparing our deployment to the GMLP principles. The purpose of this manuscript is to provide concrete insights into the deployment of ML in a healthcare setting and highlight opportunities to strengthen GMLP guidance.

Materials and methods

Model development

We developed a model to detect inpatient deterioration, defined as in-hospital death or transfer to the intensive care unit (ICU).

We obtained historical development data through the hospital's enterprise data warehouse. We used the following data sources: demographic information (sex and age), laboratory and vitals measurements. Our dataset consisted of all complete inpatient admissions to the GIM service between the dates of April 1, 2011 and December 11, 2019. We split the data into training and validation based on calendar date. Then, following silent deployment, we used data generated in the production environment between January 1, 2020 to May 30, 2020 as our test dataset. In the training and validation sets, we excluded any visits with length of stay less than 8 h or more than 40. The exclusion criteria were not applied to the test dataset. This was done to avoid biasing model development with outliers but to ensure accurate reporting of expected performance in a production environment. During training, death on the ward, transfer to the ICU, transfer to the step-up unit (a 4-bed unit on the GIM ward for higher acuity patients), and transfer to palliative care were used as the outcomes. Model performance was ultimately evaluated on the composite outcome of ICU transfer and in-hospital mortality.

We processed the data into a timeseries of 6-hour windows for each patient encounter, from admission timestamp to the first of either discharge timestamp or outcome timestamp. We took the mean value of numeric features (laboratory and vitals measurements) when the data were recorded multiple times within the same interval.

We observed a few data quality issues caused by data entry errors. For example, we found a body temperature value of 700 °C (see **Table 1** for more examples of laboratory and vitals measurements). To address this, we processed all of our numeric features in the following way: we trimmed numeric features that were less than the 1st percentile or greater than the 99th percentile (as determined from the training data), and normalized the values to be between 0 and 1, using the 1st and 99th percentile. We then created "[feature]_measured" variables to indicate whether the feature was measured in the 6-hour window and "[feature]_time_since_last_measured" variables to keep track of the number of hours since the feature was previously measured. To address missingness, we imputed data with the last observation carried forward, followed by mean imputation. Details on data processing can be found in (10).

Following numerous experiments with various machine learning methods, including logistic regression, Lasso regression, generalized additive models, and neural networks (10), we trained a "time-aware MARS model" to predict patient deterioration. This model consisted of two components: (1) The Multivariate Adaptive Regression Spline (MARS) used all processed input features to get a score of patient deterioration for each 6-hour window. (2) Then, additional features were created from the MARS scores (for example, current MARS score, baseline MARS score, change in score from previous time window, change since baseline) and were given as input to a logistic regression model. We selected this approach because it achieves similar performance to the more complex ML models, intrinsically incorporates a degree of feature selection, successfully models non-linear interactions, and was computationally efficient for deployment.

The risk scores generated by the time-aware MARS model were then categorized into "High risk", "Medium risk", and "Low risk". We used 10-fold cross-validation on the validation set to pick the thresholds that would yield a visit-level positive predictive value (PPV) of 40% and a negative predictive value (NPV) of 99%. This threshold was selected because clinicians expressed the need to minimize false alerts, and they recommended a ratio of 2 false alerts to a single true positive (10). Models were not selected or adjusted after assessing performance in the validation dataset.

Description of system

CHARTwatch was developed using the open source programming language R and the codebase was deployed to a local server with access to the hospital source systems. Various automated scripts are scheduled to run at different intervals. A summary of the scripts is provided in **Appendix Table 1**. The main CHARTwatch pipeline script runs hourly. It connects to the hospital databases, extracts the current patient census, and then pulls demographics, labs, and vitals data for current GIM patients. The script then does data cleaning, data processing, model prediction, and risk group assignment. The outputs of the CHARTwatch pipeline script are then communicated to clinicians through different methods, which are part of a comprehensive clinical intervention that was designed by an interdisciplinary team (the team has previously been described) (2):

- "High risk" alerts are sent in real-time to the GIM physicians through a paging application—"SPOK"[1]—running on the GIM team phones and the charge nurse phone. At our hospital, each GIM team and on-shift charge nurses carry a hospital-assigned mobile phone device 24-hours per day. Typically, the GIM team phones are carried by in-house residents.
- Patient risk groups are displayed, and updated hourly, in a locally-developed "electronic sign out" tool, which is used by GIM physicians to organize their teams.
- Emails are sent twice a day to the GIM charge nurses. This email contains the census of all GIM patients and their CHARTwatch risk group. The email is used by the charge nurses when they are assigning bedside nurses for the next 12-hour shift. They proactively attempt to match more experienced or skilled nurses with higher risk patients and to avoid one nurse from having multiple "High risk" patients.
- An email is sent once daily to the Palliative care team. This email contains a list of all patients who received their first "High risk" prediction in the past 24 h. The palliative care team contacts the patients' GIM physicians to ask whether a palliative care consultation would be helpful, with the goal of improving access to high quality end-of-life care for "High risk" patients, when appropriate.

[1] https://www.spok.com/

"High risk" alerts to the mobile devices are triggered if a patient is classified as "High risk" by the CHARTwatch model. To minimize alert fatigue, the following alerting protocol is applied:

- After an alert, no further alerts are sent for the same patient for the next 48 h.
- Subsequent alerts only occur if a patient's status changes from "High risk" to "Low risk" or "Medium risk", and then back to "High risk".
- Alerts for patients who are transferred from the ICU to the GIM ward are silenced for 24 h after transfer, as these patients are already known to be "High risk" and are proactively followed by the critical care response team.
- In March 2022, in response to feedback that a small number of patients were getting a very large number of alerts, we began silencing all alerts after the fifth alert for a patient, although their status is still indicated as "High risk" in other communications.

A clinical pathway for "High risk" alerted patients was developed by an implementation team as described in detail elsewhere (2). This pathway was continuously refined by an implementation committee composed of GIM, ICU, and palliative care physicians and nurses, the chief medical resident, a clinical informatics specialist, and the lead data scientist. The committee met weekly through the pilot and early phases of the implementation and then was scaled back to meet monthly.

Silent deployment and pilot phase

CHARTwatch silent deployment was affected by the onset of the COVID-19 pandemic. We initially planned a 4-month silent deployment, which was subsequently extended for 6 more months to accommodate the clinical changes that were being made amidst the challenges of the first wave of the pandemic. During silent testing, we used weekly check-in meetings with stakeholders to ensure the system was running smoothly and had some preliminary training sessions with end-users to assess buy-in and trust. The weekly check-in meetings included members from the following groups: the model development team, hospital Information Technology (IT), clinical informatics, and clinicians working on GIM, ICU and Palliative Care.

We began silent deployment in November 2019 and planned to launch the intervention in March 2020. During this time, we focused on several data-related issues. First, our testing period coincided with the hospital changing from traditional troponin measurement to a new "high sensitivity" troponin assay. In order to address this change, we modified our pre-processing code to scale the lab value accordingly. Earlier versions of CHARTwatch relied on medications and nurse notes. However, this silent testing period uncovered database connection issues, and these data sources were subsequently removed. This had no impact on model performance.

The onset of the COVID-19 pandemic resulted in numerous changes on the GIM unit, as this was the unit primarily responsible for care of COVID-19 ward patients. This resulted in physical changes to the ward, relocation of patients to other units in the hospital, and creation of new clinical care teams. All of these changes needed to be accommodated in CHARTwatch, including ensuring the model continued to identify the correct cohort of GIM patients and would be delivered to all the relevant care teams. We made a plan to deploy the model for GIM patients with COVID-19 as well as for those with other illnesses. Once these changes were made, we focused on model validation and data quality to ensure accurate performance (see Results section for details).

In August 2020, we initiated the pilot phase of our intervention. We began by deploying the system for 2 of the 7 GIM teams. Deployment progressed in a phased approach over a 6-week period, rolling out to all GIM teams and then to nurses and the palliative care team. The system was fully running by October 20, 2020.

The silent deployment and pilot phases were essential as it allowed the technical team to uncover issues with pipelines and workflows, and also allowed the clinical team to collect feedback from end-users.

A summary of major changes resulting from silent deployment and iterative refinement of the solution during the pilot and implementation phases is provided in **Appendix Table 2**.

Downtime protocols

To ensure that CHARTwatch could run smoothly with minimal interruptions, all deployed scripts were developed so that emails to the project team would be sent out if any script failed. Furthermore, the data extract scripts were set to run hourly. If data extraction failed, the model outputs could rely on an earlier data extract that is at most 3 h old. In the rare case where errors would affect end-users, we developed email templates to: (1) notify end-users of the unplanned downtime; and (2) notify them when the unplanned downtime was over.

Planned downtimes are inevitable (e.g., due to database updates, server updates) and we also developed email templates accordingly.

End-user engagement

End-users were engaged through the full life-cycle of this project as described in the methods and previous manuscripts

(2, 10). A comprehensive effort was made to train physicians and nurses prior to, and during the deployment of CHARTwatch about the system, the interpretation of risk groups, the meaning of alerts, and the expected clinical responses. We incorporated CHARTwatch training into the orientation of all new nurses, such that all nurses working on the GIM ward receive CHARTwatch training. CHARTwatch training was integrated into the monthly orientation for residents, which includes in-person and emailed materials, and involves approximately 100 resident physicians annually. GIM staff physicians were trained through several presentations at division meetings and all GIM physicians (approximately 20) received the training.

Performance measures

To measure model performance, we report the area under the receiver-operative curve (AUC), the positive prediction value (PPV), and sensitivity. We compute these metrics in the heldout test data (January 2020–May 2020), and in the real-time data from deployment (August 2020–March 2022). The model PPV is computed at the encounter-level. We calculate sensitivity based on the visit's maximum risk group, in order to get an estimate of the proportion of outcomes that would be captured by the visit's maximum risk group. We want this metric to be low for patients whose maximum risk group is “Low risk”.

To estimate adherence to clinical pathways, we used the number of vital sign assessments in the 24 h following an alert. This reflects both physician and nurse adherence as physicians must place an order and nurses must perform the measurements. According to the clinical pathway, vital signs should be measured every 4 h (the maximum frequency of routine measurement for patients on the GIM ward). Thus, adherence is measured as the total number of alarms that follow the clinical pathway (i.e., are followed by vital signs measurements every four hours) divided by the total number of alarms. We compute this metric at a weekly level and report the weekly percentage of alerts that follow the clinical pathway.

Results

Model performance

Model performance metrics on the heldout test data (January 2020 to May 2020, silent testing period) and the deployment data (August 2020 to March 2022) are reported in **Table 2**. When predicting the composite outcome of ICU transfer and in-hospital mortality, the time-aware MARS model achieved an AUC of 0.786 and of 0.759 on the test data and deployment data, respectively. When predicting patient deterioration within the next 48 h, the AUC was of 0.626 and 0.753.

Maximum risk group sensitivity was of 0.530 for the “High risk” group, 0.471 for the “Medium risk” group, and 0 for the “Low risk” group in the test data. In production, the risk group sensitivity was of 0.559, 0.417, and 0.023.

During validation, we iterated through a range of risk score values and selected a threshold that would yield a PPV of 0.40 on the composite outcome of ICU transfer, in-hospital mortality, step-up unit transfer, and Palliative Care transfer. With this outcome, the model achieved a PPV of 0.306 and 0.272 in the test data and deployment data, respectively. In the composite outcome of ICU transfer and in-hospital mortality only, the model achieved a PPV of 0.172 and 0.257 in the test and deployment data, respectively.

Alerts

Since deployment, there has been a mean of 2.60 (SD: 1.71) alerts per day and a median of 2 (IQR: [1–4]), for a mean of 84.02 (SD: 7.48) and median 84 (IQR: [79–89]) total GIM patients per day. There were 56 (9.589%) days where no alerts were sent. **Figure 1** shows the daily number of alerts since deployment. The alerts were equally spread out across the different GIM teams.

Adherence to clinical pathway

To assess adherence, we reported the percentage of alerted patients who had at least 4 vital signs measurements. The weekly percentage of adherence increased as users became more familiar with the system. Between August 2020 and November 2020, this weekly percentage was at a mean of 65% (SD: 11%) and a median of 65% (IQR: [59%–75%]). Between December 2021 and March 2022, this weekly rate had increased to a mean of 74% (SD: 11%) and a median of 71% (IQR: [69%–80%]).

CHARTwatch pipeline runtime and data size

The pipeline runtime remains consistent and takes a mean of 196.83 seconds (SD: 121.05 seconds) to complete and a median of 151 seconds (IQR: [133–194] seconds). Similarly, the data size remains consistent at a mean of 26.44 MB (SD: 3.98 MB) and a median of 26.25 MB (IQR: [24.42–28.06] MB).

Downtime events

We experienced few downtime events and most of them were planned. In total, the system was down for 52.5 out of 14,016 h (584 days). Thus, CHARTwatch was running for 99.6% of the time. 20 h (38.1%) of downtime were planned (scheduled database maintenance/upgrade, updates to server, etc.), and 32.5 h (61.9%) of downtime were due to unplanned events (such as unexpected database or network failure).

Adherence to GMLP guidelines

1. **Multi-Disciplinary Expertise Is Leveraged Throughout the Total Product Life Cycle**. CHARTwatch was developed and deployed by a team from various fields of expertise with strong end-user engagement including advice from patients and caregivers, as previously described in detail (2).
2. **Good Software Engineering and Security Practices Are Implemented**. Our infrastructure follows best practices for security; the deployment server for CHARTWatch sits in the same secure private network as the clinical systems. Access to all systems is restricted to authorized personnel and continuously audited. Database administrators of clinical systems provided guidance to data engineers in developing high performance queries. The data pipelines were coded using techniques to minimize the risk of SQL injection in case of a system breach, while leaving a minimal footprint on the source systems. Furthermore, the data science team employed an agile development approach to develop the final deployed product. This included regular meetings to assess tasks, re-visiting the backlog and prioritizing as needed.
3. **Clinical Study Participants and Data Sets Are Representative of the Intended Patient Population**. The model was trained on historical hospitalizations from the same patient population at the same institution, to maximize representativeness. During model development, we worked directly with source database systems to ensure high quality data, including performing clinical validation ensured that the data sets were representative of real-world data.
4. **Training Data Sets Are Independent of Test Sets.** Our training data sets and test data sets were independent of each other. We used calendar-based data split approach to ensure that performance reported on the test set would be representative of deployment-level performance by simulating historical training and deployment in a future time period. Furthermore, silent testing did not overlap with our training/validation/test datasets.
5. **Selected Reference Datasets Are Based Upon Best Available Methods.** We tried multiple models and settled on the one that gave us the best performance and could work within constraints and limitations set by the system. Further, the features and models used in CHARTwatch were backed by previous evidence (10, 11).
6. **Model Design Is Tailored to the Available Data and Reflects the Intended Use of the Device**. When developing the model, we used data available at time of each prediction. Any data generated or updated after the expected prediction time was excluded from the training dataset. The silent deployment periods also allowed us to validate this.
7. **Focus Is Placed on the Performance of the Human-AI Team.** The clinical team was involved in development and deployment with regular meetings with all stakeholders and with extensive training on how to use the system. We suggested CHARTwatch predictions be used by clinicians in conjunction with their own clinical judgement rather than in isolation. Further, we conducted a clinical validation, comparing CHARTwatch model predictions to more than 3,000 real-time clinical predictions, to engender trust and inform our understanding of the human-AI team (manuscript under review).
8. **Testing Demonstrates Device Performance during Clinically Relevant Conditions**. We had silent testing periods as well as a pilot phase and a phased rollout. Model performance was monitored throughout the silent testing period and continues to be monitored on an ongoing basis.

TABLE 2 Performance of the CHARTwatch model in the test data and the deployment data. AUC, PPV, and sensitivity are reported in the test data (January 2020–May 2020) and in the deployment data (August 2020–March 2022). Metrics are reported on the composite outcome of ICU transfer and in-hospital mortality (Outcome: ICU/Death), as well as in the composite outcome of ICU transfer, in-hospital mortality, step-up unit transfer, and Palliative Care transfer (Outcome: ICU/Death/step-up/Palliative).

Metric	Test Data		Deployment Data	
	Outcome: ICU/ Death	Outcome: ICU/ Death/ step-up/ Palliative	Outcome: ICU/ Death	Outcome: ICU/ Death/ step-up/ Palliative
AUC (ever)	0.786	0.735	0.759	0.768
AUC (in next 48 h)	0.626	0.791	0.753	0.759
PPV of alerted encounters	0.172	0.306	0.257	0.272
Sensitivity (based on maximum risk group)				
High risk	0.480	0.53	0.565	0.559
Medium risk	0.520	0.471	0.419	0.417
Low risk	0	0	0.016	0.023

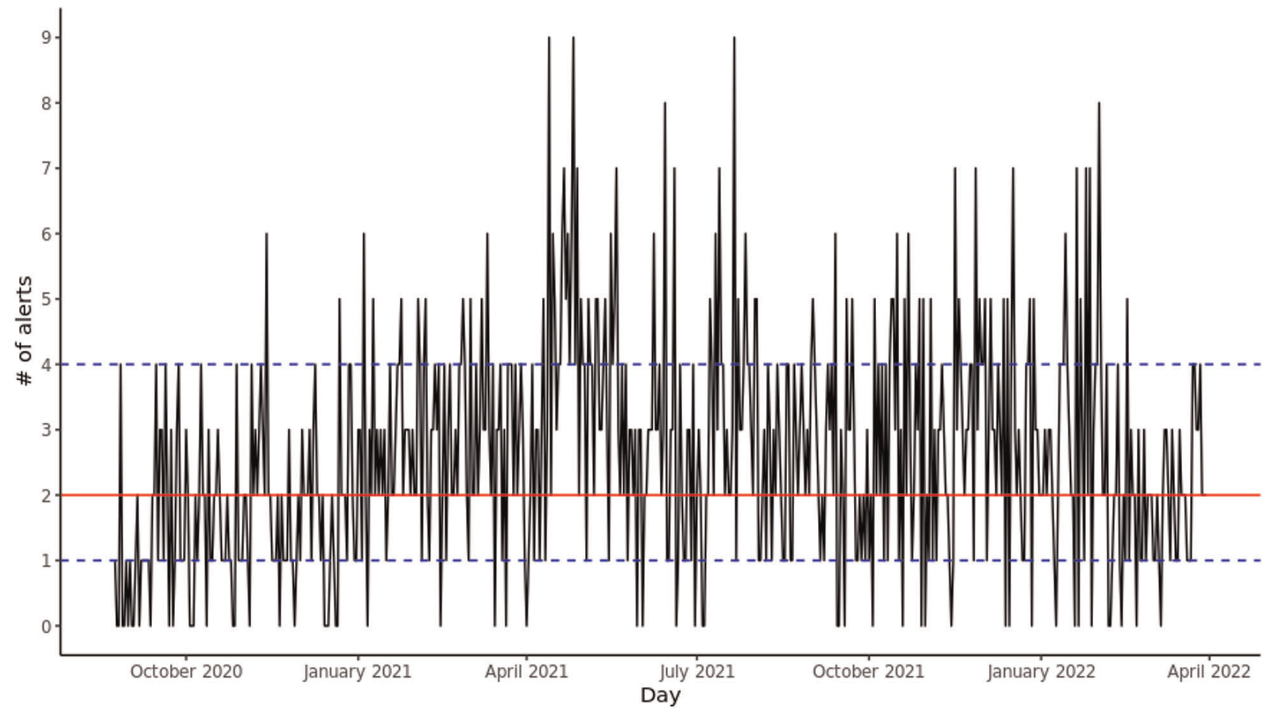


FIGURE 1
Daily alerts sent by CHARTwatch. The red solid line indicates the median number of daily alerts. The blue dashed lines indicate the 25th quantile and the 75th quantile.

9. **Users Are Provided Clear, Essential Information**. To ensure the delivery of simple, actionable messages for clinicians, CHARTwatch predictions were categorized into "High risk", "Medium risk", and "Low risk" groups. Messaging alerts contain the following text: "[Patient Last Name, First Name, Medical Record Number] is high risk for transfer to ICU or death. Please refer to **LINK** for more information." The link takes users to a brief description of the clinical protocol for alerted patients. All clinicians receive training on how to use this system.
10. **Deployed Models Are Monitored for Performance and Re-training Risks are Managed.** Model performance is measured and monitored by an implementation committee, using a small number of key performance and process measures, including those reported in this manuscript: model sensitivity, PPV, number of daily alerts, number of outcomes for patients in different risk groups, and number of vital signs measurements in the 24 h following an alert. These were initially monitored weekly and once the intervention moved into a more stable maintenance phase, committee meetings are held monthly. Error alerts on the automated pipelines ensure timely identification of errors by the team. Re-training poses an important challenge, as the model has altered clinical workflows, particularly for alerted patients. Given that clinical interventions are intended to prevent adverse outcomes for alerted patients, re-training the model may lead to undesired feedback loops resulting in poorer performance. This remains an area of active research for our team and others (12), including exploring the use of proxy labels to ensure that high risk patients who do not experience adverse outcomes are still captured in the modelling.

Discussion

In this manuscript, we describe our experience deploying an early warning system for GIM patients in an academic hospital, which highlights numerous practical lessons. We observed that the GMLP guiding principles offer a helpful starting point, and our solution was developed in alignment with these suggestions. We offer concrete and detailed descriptions of how we were able to operationalize the various GMLP recommendations, to assist future initiatives. Beyond these principles, we identified several aspects that have been critical for successful deployment of our solution. First, engagement of end-users was essential in designing, deploying, and iteratively refining the solution. Second, a silent testing period and phased launch was crucial for identifying unanticipated issues with models and data pipelines and resulted in numerous updates before launch. Third, it was important to create robust downtime protocols, with a careful plan to prevent disruptions in clinical workflow or patient harm.

Engaging a multidisciplinary group of end-users from the project's outset ensured that there was a high level of trust and uptake of the designed solution. We discuss in more detail our findings of engaging a multidisciplinary group in (2). Ongoing engagement led to important iterations in the intervention. Our engagement included identifying key champions to participate in committees and lead the initiative and a comprehensive training program for all clinicians. Regular implementation committee meetings, initially held weekly and then scaled back over time to monthly, allowed the team to refine the intervention in response to feedback from clinicians.

We achieve high model performance and, on average, CHARTwatch only sends out two alerts each day. The outcomes not captured by the alert are captured by the "Medium risk" group. Combining the "High risk" and "Medium risk" groups together yields a sensitivity of 0.976. In addition to monitoring various measures of model performance and clinical outcomes, we used a simple process measure to capture clinical adherence (the number of vital signs measurements). This reflects both physician and nurse practice and demonstrated good adherence at the project's outset, with further improvements over time.

There is a notable absence of "best practices" in deploying ML solutions in healthcare. The GMLP guiding principles are an important step forward and as high-level guides, they are very useful. However, greater specificity is needed to understand how these principles can be operationalized, and this manuscript reflects an effort to provide some of that additional detail. We also note that there are several crucial areas for ML deployment which are referred to only tangentially in the GMLP and ought to be mentioned specifically. GMLP Principle 1 ("Multi-Disciplinary Expertise Is Leveraged Throughout the Total Product Life Cycle") is very applicable to the importance of end-user engagement, although it does not mention this specifically. GMLP Principle 8 ("Testing Demonstrates Device Performance during Clinically Relevant Conditions") may be strengthened by highlighting the importance of silent testing in a real-time production environment before deployment into clinical care. An important area for future research is to develop a guiding framework that would help determine what duration of silent testing is sufficient before deployment. This duration would be affected by parameters related to model performance (e.g., prevalence of outcome events, desired model accuracy) and factors related to the data pipeline and clinical context (e.g., number of clinical systems involved, planned updates to systems and processes). GMLP Principles 9 ("Users Are Provided Clear, Essential Information") and 10 ("Deployed Models Are Monitored for Performance and Re-training Risks are Managed") should be expanded to include downtime protocols. While system and model failures are rare, they are bound to happen, and end-users should not be left in the dark. Borrowing from safety engineering, failure modes and effect analysis could be a good way to identify all potential risks within the deployed system and, accordingly, develop downtime protocols (13). Finally, we note that GMLP Principle 10 raises a crucial area for future research. Re-training models that have already been deployed into, and affected, clinical practice raises challenging methodological issues (12). Identifying methods to maintain highly accurate models over time is an urgent need as models are increasingly deployed into clinical environments.

This work has several limitations. First, our deployment was conducted in a single academic hospital and thus generalizability to other settings must be considered. However, we believe the key lessons from our experience are very likely to apply to a wide range of ML solutions. Second, our deployment relates to a clinical decision-support and predictive analytics solution. Other ML applications (e.g., computer vision) may require a different set of approaches for their deployment. Third, we relied primarily on routinely-collected data to measure model performance and clinical adherence. This has the advantage of being scalable and resource-efficient, but lacks granularity and clinical context. Targeted chart reviews, as have been described in the quality improvement literature (14), or interviews with clinicians represent other important ways of gathering this information.

In conclusion, deploying machine learning models in healthcare settings is challenging and requires a multidisciplinary team to ensure success. As these deployments become more frequent, we hope that more rigorous standards and best practices will arise. The evolution of the GMLP guiding principles, and lessons learned from real-world implementations, can assist with strengthening best practices in the deployment of machine learning models.

Data availability statement

The data is not publicly available as it contains personal health information (PHI).

Ethics statement

The study was reviewed and approved by the St. Michael's Hospital Research Ethics Board. Patient consent was not required as the deployment of the AI system is not a research intervention, but rather a hospital supported quality improvement intervention.

Author contributions

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. All authors contributed to the article and approved the submitted version.

Funding

This project received funding from the Pathfinder Project grant from the Vector Institute and from the Li ka Shing Knowledge Institute of St. Michael's Hospital.

Conflict of interest

All authors hold shares in Signal 1, a health artificial intelligence solutions company.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/https://www.frontiersin.org/articles/10.3389/fdgth.2022.932123/full#supplementary-material.

References

1. Ben-Israel D, Jacobs WB, Casha S, Lang S, Ryu WHA, de Lotbiniere-Bassett M, et al. The impact of machine learning on patient care: A systematic review. *Artif Intell Med.* (2020) 103:101785. doi: 10.1016/j.artmed.2019.101785

2. Verma AA, Murray J, Greiner R, Cohen JP, Shojania KG, Ghassemi M, et al. Implementing machine learning in medicine. *CMAJ.* (2021) 193(34):E1351–7. doi: 10.1503/cmaj.202434

3. Drysdale E, Dolatabadi E, Chivers C, Liu V, Saria S, Sendak M, et al. *Implementing AI in healthcare.* Toronto, ON: Vector-SickKids Health AI Deployment Symposium (2020). 27 p.

4. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: A roadmap for responsible machine learning for health care. *Nat Med.* (2019) 25(9):1337–40. doi: 10.1038/s41591-019-0548-6

5. Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med.* (2018) 178(11):1544–7. doi: 10.1001/jamainternmed.2018.3763

6. Cohen JP, Cao T, Viviano JD, Huang CW, Fralick M, Ghassemi M, et al. Problems in the deployment of machine-learned models in health care. *CMAJ.* (2021) 193(35):E1391–4. doi: 10.1503/cmaj.202066

7. Paulson SS, Dummett BA, Green J, Scruth E, Reyes V, Escobar GJ. What do we do after the pilot is done? Implementation of a hospital early warning system at scale. *Jt Comm J Qual Patient Saf.* (2020) 46(4):207–16. doi: 10.1016/j.jcjq.2020.01.003

8. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé III H, et al. Datasheets for datasets. *Commun ACM.* (2021) 64(12):86–92. doi: 10.1145/3458723

9. U.S. Food & Drug Administration. Good machine learning practice for medical device development: Guiding principles. FDA (2021). Available from: https://www.fda.gov/medical-devices/software-medical-device-samd/good-machine-learning-practice-medical-device-development-guiding-principles (cited 2021 Dec 8).

10. Nestor B, McCoy LG, Verma A, Pou-Prom C, Murray J, Kuzulugil S, 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). p. 950–72. Available at: https://proceedings.mlr.press/v126/nestor20a.html (cited 2022 Jan 20).

11. Suresh H, Hunt N, Johnson A, Celi LA, Szolovits P, Ghassemi M. Clinical intervention prediction and understanding using deep networks. ArXiv170508498 Cs (2017). Available at: http://arxiv.org/abs/1705.08498 (cited 2021 Oct 4).

12. Adam GA, Chang CHK, Haibe-Kains B, Goldenberg A. *Hidden risks of machine learning applied to healthcare: unintended feedback loops between models and future data causing model degradation. Proceedings of the 5th Machine Learning for Healthcare Conference.* PMLR (2020). p. 710–31. Available at: https://proceedings.mlr.press/v126/adam20a.html (cited 2022 Apr 27).

13. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, et al. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. ArXiv200100973 Cs(2020). Available at: http://arxiv.org/abs/2001.00973 (cited 2022 Feb 16).

14. Etchells E, Woodcock T. Value of small sample sizes in rapid-cycle quality improvement projects 2: Assessing fidelity of implementation for improvement interventions. *BMJ Qual Saf.* (2018) 27(1):61–5. doi: 10.1136/bmjqs-2017-006963

15. Friedman JH. Multivariate adaptive regression splines. *Ann Stat.* (1991) 19(1):1–67. doi: 10.1214/aos/1176347963

16. R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing (2020). Available from: https://www.R-project.org/

17. Tibshirani SMD from mda:mars by TH and R. earth: Multivariate Adaptive Regression Splines [Internet]. 2011. Available at: http://CRAN.R-project.org/package=earth

18. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* (2010) 33(1):1–22. doi: 10.18637/jss.v033.i01

19. Kuhn M, Wickham H. Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles (2020). Available at: https://www.tidymodels.org

Appendix Table 1: Table of scheduled scripts.

Script	Description	Schedule
CHARTwatch pipeline	The script extracts data from the source systems, processes the data, generates model prediction, and classifies each patient into risk groups (“Low risk”, “Medium risk”, “High risk”).	Hourly
Charge nurse email	The scripts sends a list of the patient census, including CHARTwatch risk groups, to the charge nurse twice a day.	Every 12 h
SPOK alert update	The script sends alerts to the “SPOK” application on the GIM team phones and charge nurse phone. The script sends alerts on “High risk” patients and applies re-alerting silencing rules (as specified in Section “Description of system”).	Hourly
“Electronic sign out” tool update	The script updates the CHARTwatch risk groups in the “electronic sign out” database.	Hourly
Palliative team email	The script sends a daily email to the Palliative care team. The email contains a list of all new “High risk” patients.	Daily

Appendix Table 2: Timeline of CHARTwatch major changes

Date	Change	Type of change
13-Nov-19	**Silent deployment**. Deployment was on an older server and consisted in an ensemble model that used the following input data: demographics, labs, vitals, medications, clinical orders, nursing notes.	Silent deployment
26-Nov-19	New “high sensitivity troponin” lab added.	Process change
20-Dec-19	**Silent deployment update.** The model was updated to an ensemble that relied only on demographics, labs, and vitals. We also fixed an issue where sodium labs (lab code is “NA”) were getting interpreted as “Not available”.	Silent deployment
24-Dec-19	**Silent deployment update.** The data processing was changed to address the troponin lab update.	Silent deployment
14-May-20	**Silent deployment update**. The deployment was moved to a different server and we used the time-aware MARS model (as described in Section “Model development”).	Silent deployment
25-Aug-20	**Start of pilot phase.** Alerts and ‘electronic sign out’ was activated for 2 GIM teams.	Deployment
11-Sep-20	Risk group rule change: if patient is on step-up unit, their risk group must at minimum be “Medium risk”.	CW change
11-Sep-20	Alerting rule change: Alerts silenced for 24 h after patient leaves ICU.	CW change
11-Sep-20	Add “Team Stroke” to data extraction query.	Process change
11-Sep-20	Add new GIM ward location to data extraction query, corresponding to the opening of a new patient care tower.	Process change
15-Sep-20	**Launch to all GIM teams.**	Deployment
6-Oct-20	**Deployment to charge nurses.** Emails are sent to the charge nurse email address twice a day. Alerts are sent to the charge nurse phone.	Deployment
20-Oct-20	**Deployment to palliative team—Full deployment.** Emails are sent daily to the Palliative Care team email.	Deployment
19-Jan-21	Switch from alerts 3×/day to hourly alerts.	CW change
27-Apr-21	Add an extra GIM team (opened for COVID-19) to data extraction query.	Deployment
11-Jun-21	Remove extra COVID-19 team from data extraction query, as team closed.	Deployment
8-Mar-22	Alerting rule change: stop repeat alerts after 5th alert.	CW change

Appendix 3: Details on the CHARTwatch model

The final deployed model consists of two pieces:

(1) The Multivariate Adaptive Regression Spline (MARS) is a weighted sum of basis functions (15). To determine the coefficients, we used cross-validation to train the model. The inputs to the MARS model are the features described in Section "Model development".

(2) The output of the MARS model is a score ranging from 0 to 1. The MARS score, as well as the difference in MARS score between this time window and the previous one, the percent change, and the percent change from baseline, are then given as input to a logistic regression model. The logistic model is trained on the training data.

The MARS and logistic regression models were trained using R (version 3.6.3) (16) with the packages 'earth' (version 5.1.2) (17) and 'glmnet' (4.0-2), respectively (18). We use the 'tidymodels' suite of packages (19) to train and tune the models.

frontiers | Frontiers in Digital Health

TYPE Original Research
PUBLISHED 12 September 2022
DOI 10.3389/fdgth.2022.943768

Considerations in the reliability and fairness audits of predictive models for advance care planning

Jonathan Lu[1]*, Amelia Sattler[2†], Samantha Wang[3†], Ali Raza Khaki[4†], Alison Callahan[1], Scott Fleming[1], Rebecca Fong[5], Benjamin Ehlert[1], Ron C. Li[3], Lisa Shieh[3], Kavitha Ramchandran[4], Michael F. Gensheimer[6], Sarah Chobot[7], Stephen Pfohl[1], Siyun Li[1], Kenny Shum[8], Nitin Parikh[8], Priya Desai[8], Briththa Seevaratnam[5], Melanie Hanson[5], Margaret Smith[2], Yizhe Xu[1], Arjun Gokhale[1], Steven Lin[2], Michael A. Pfeffer[3,8], Winifred Teuteberg[5‡] and Nigam H. Shah[1,8,9‡]

[1]Center for Biomedical Informatics Research, Department of Medicine, Stanford University School of Medicine, Palo Alto, United States, [2]Stanford Healthcare AI Applied Research Team, Division of Primary Care and Population Health, Department of Medicine, Stanford University School of Medicine, Palo Alto, United States, [3]Division of Hospital Medicine, Department of Medicine, Stanford University School of Medicine, Palo Alto, United States, [4]Division of Oncology, Department of Medicine, Stanford University School of Medicine, Palo Alto, United States, [5]Serious Illness Care Program, Department of Medicine, Stanford University School of Medicine, Palo Alto, United States, [6]Department of Radiation Oncology, Stanford University School of Medicine, Palo Alto, United States, [7]Inpatient Palliative Care, Stanford Health Care, Palo Alto, United States, [8]Technology & Digital Solutions, Stanford Health Care and Stanford University School of Medicine, Palo Alto, United States, [9]Clinical Excellence Research Center, Stanford University School of Medicine, Palo Alto, United States

Multiple reporting guidelines for artificial intelligence (AI) models in healthcare recommend that models be audited for reliability and fairness. However, there is a gap of operational guidance for performing reliability and fairness audits in practice. Following guideline recommendations, we conducted a reliability audit of two models based on model performance and calibration as well as a fairness audit based on summary statistics, subgroup performance and subgroup calibration. We assessed the Epic End-of-Life (EOL) Index model and an internally developed Stanford Hospital Medicine (HM) Advance Care Planning (ACP) model in 3 practice settings: Primary Care, Inpatient Oncology and Hospital Medicine, using clinicians' answers to the surprise question ("Would you be surprised if [patient X] passed away in [Y years]?") as a surrogate outcome. For performance, the models had positive predictive value (PPV) at or above 0.76 in all settings. In Hospital Medicine and Inpatient Oncology, the Stanford HM ACP model had higher sensitivity (0.69, 0.89 respectively) than the EOL model (0.20, 0.27), and better calibration (O/E 1.5, 1.7) than the EOL model (O/E 2.5, 3.0). The Epic EOL model flagged fewer patients (11%, 21% respectively) than the Stanford HM ACP model (38%, 75%). There were no differences in performance and calibration by sex. Both models had lower sensitivity in Hispanic/Latino male patients with Race listed as "Other." 10 clinicians were surveyed after a presentation summarizing the audit. 10/10 reported that summary statistics, overall performance, and subgroup performance would affect their decision to use

Check for updates

OPEN ACCESS

EDITED BY
Mark Sendak,
Duke University School of Medicine,
United States

REVIEWED BY
Ahmet Yardimci,
Akdeniz University, Turkey
Matloob Khushi,
University of Suffolk, United Kingdom

*CORRESPONDENCE
Jonathan Hsijing Lu
jhlu@stanford.edu

†These authors have contributed equally to this work.

‡These authors have contributed equally to this work and share senior authorship.

SPECIALTY SECTION
This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health

RECEIVED 14 May 2022
ACCEPTED 17 August 2022
PUBLISHED 12 September 2022

CITATION
Lu J, Sattler A, Wang S, Khaki AR, Callahan A, Fleming S, Fong R, Ehlert B, Li RC, Shieh L, Ramchandran K, Gensheimer MF, Chobot S, Pfohl S, Li S, Shum K, Parikh N, Desai P, Seevaratnam B, Hanson M, Smith M, Xu Y, Gokhale A, Lin S, Pfeffer MA, Teuteberg W and Shah NH (2022) Considerations in the reliability and fairness audits of predictive models for advance care planning.
Front. Digit. Health 4:943768.
doi: 10.3389/fdgth.2022.943768

the model to guide care; 9/10 said the same for overall and subgroup calibration. The most commonly identified barriers for routinely conducting such reliability and fairness audits were poor demographic data quality and lack of data access. This audit required 115 person-hours across 8–10 months. Our recommendations for performing reliability and fairness audits include verifying data validity, analyzing model performance on intersectional subgroups, and collecting clinician-patient linkages as necessary for label generation by clinicians. Those responsible for AI models should require such audits before model deployment and mediate between model auditors and impacted stakeholders.

KEYWORDS

model reporting guideline, electronic health record, artificial intelligence, advance care planning, fairness, audit

Introduction

Concern about the reliability and fairness of deployed artificial intelligence (AI) models trained on electronic health record (EHR) data is growing. EHR-based AI models have been found to be unreliable, with decreased performance and calibration across different geographic locations and over time; for example, an Epic sepsis prediction algorithm had reduced performance when validated by University of Michigan researchers (1) and acute kidney injury models have shown worsening calibration over time (2). AI models have also been found to be unfair, with worse performance and calibration for historically marginalized subgroups; for example, widely used facial recognition algorithms have lower performance on darker-skinned females (3); and widely used health insurance algorithms underrate the disease status of Black patients compared with similar White patients (4). Despite lacking evidence of reliability and fairness, algorithms are still being deployed (5).

To promote improved reliability and fairness of deployed EHR models, at least 15 different model reporting guidelines have been published (6–20). Some commonly included items related to reliability in these guidelines include external validation (6, 8–10, 14–17, 19); multiple performance metrics such as Area Under Receiver Operating Curve (AUROC) (6, 8–12, 14–18), positive predictive value (PPV) (9–12, 14, 16–18), sensitivity (8–12, 14, 16–18), and specificity (8–12, 14, 17, 18); confidence intervals or another measure of variability of the performance (6, 8–12, 15, 18–20); and calibration plots (6, 8–10, 12, 14). Some commonly included items related to fairness include summary statistics (10, 11, 15, 17, 18, 20), like the distribution of demographics such as sex (11, 15, 17, 20) and race/ethnicity (15, 17, 20), as well as subgroup analyses that investigate how a model performs for specific subpopulations (7, 9, 11–13, 15, 18, 20). Nevertheless, many of these items are infrequently reported for both published (21) and deployed EHR models (22).

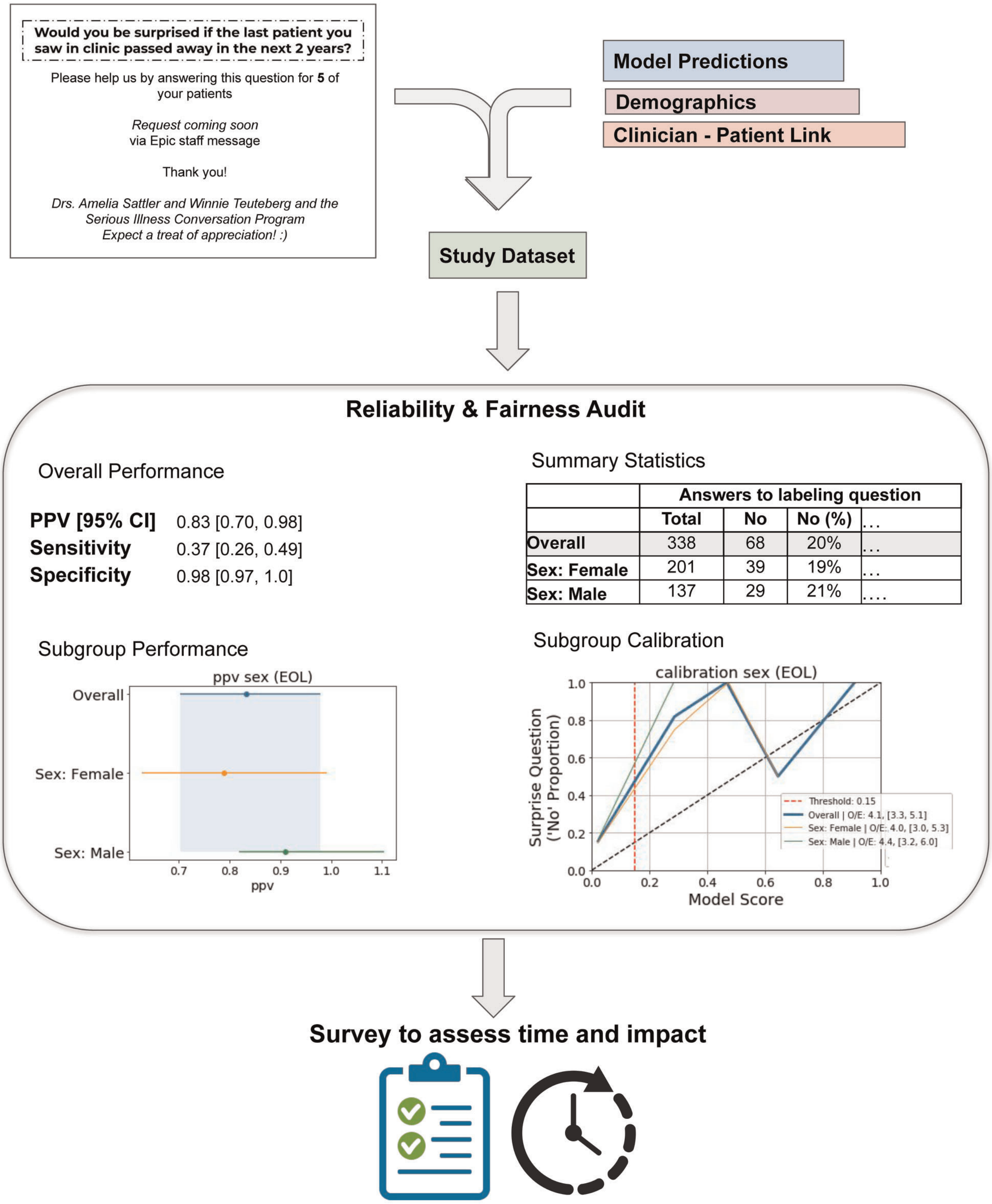
Several efforts seek to address this reporting gap. For example, there is an existing auditing framework that supports AI system development end-to-end and links development decisions to organizational values/principles (23). There is also currently an open-source effort to better understand, standardize and implement algorithmic audits (24).

In this work, we illustrate a reliability/fairness audit of 12-month mortality models considered for use in supporting team-based ACP in three practice settings (Primary Care, Inpatient Oncology, Hospital Medicine) at a quaternary academic medical center in the United States (25–27) (**Figure 1**). We (1) design and report a reliability/fairness audit of the models following existing reporting guidelines, (2) survey decision makers about how the results impacted their decision of whether to use the model, and (3) quantify the time, workflow and data requirements for performing this audit. We discuss key drivers and barriers to making these audits standard practice. We believe this may aid other decision makers and informaticists in operationalizing regular reliability and fairness audits (22, 23).

Note: we use recorded race/ethnicity in the EHR as a way to measure how models may perform across such groupings, as recommended (15, 21). *Importantly, race/ethnicity is not used as an input for any of the models and we do not use it as a "risk factor" for health disparities* (28–30). *We recognize race/ethnicity has widely varying definitions* (31) *and is more a social construct* (32) *than a biological category* (30). *We also caution that studies have found poor concordance of race/ethnicity data as recorded in the EHR with the patient's self-identification* (33, 34). *However, performance by race/ethnicity subgroups is a recommended analysis in reporting guidelines.*

Background on advance care planning and model usage

Much of care for patients at the end of their lives is not goal-concordant, i.e. not consistent with the patients' goals and values. For example, a survey (35) of Californians' attitudes towards death and dying found that 70% would prefer to die at home. Despite this, only 30% of all deaths happened at home in 2009. Meanwhile 60% occurred in a hospital or nursing home (26).



Would you be surprised if the last patient you saw in clinic passed away in the next 2 years?

Please help us by answering this question for **5** of your patients

Request coming soon via Epic staff message

Thank you!

Drs. Amelia Sattler and Winnie Teuteberg and the Serious Illness Conversation Program Expect a treat of appreciation! :)

PPV [95% CI]	0.83 [0.70, 0.98]
Sensitivity	0.37 [0.26, 0.49]
Specificity	0.98 [0.97, 1.0]

	Answers to labeling question			
	Total	No	No (%)	…
Overall	338	68	20%	…
Sex: Female	201	39	19%	…
Sex: Male	137	29	21%	….

Subgroup Performance
ppv sex (EOL)
Overall
Sex: Female
Sex: Male
0.7
0.8
0.9
1.0
1.1
ppv
Subgroup Calibration
calibration sex (EOL)
Surprise Question ('No' Proportion)
0.0
0.2
0.4
0.6
0.8
1.0
Model Score
Threshold: 0.15
Overall | O/E: 4.1, [3.3, 5.1]
Sex: Female | O/E: 4.0, [3.0, 5.3]
Sex: Male | O/E: 4.4, [3.2, 6.0]
Survey to assess time and impact

FIGURE 1
Overview of audit process. Results and plots listed are for the Epic End-of-life Index Low Threshold for Primary Care. The "labeling question" under Summary Statistics is "Would you be surprised if this patient passed away in 2 years?".

In 2018 the Stanford Department of Medicine began implementation of Ariadne Labs' Serious Illness Care Program (SICP) (36) to promote goal-concordant care by improving timing and quality of advance care planning conversations. By following best practices (37), the Stanford SICP trained and supported clinicians in using the structured Serious Illness Conversation Guide (SICG) in their practice.

Through the duration of this audit, *Primary Care* and *Inpatient Oncology* were developing implementation plans, while *Hospital Medicine* had an active implementation after SICG training of key physicians and staff members using a 12-month mortality model to generate patient prognoses that were shared with the entire clinical team (25). Two models were considered: (1) the 12-month mortality model which runs only on currently hospitalized patients only and is currently used by the Hospital Medicine SICP team (HM ACP), and (2) the Epic End-of-life (EOL) Index, which unlike HM ACP, runs for all patients receiving care in the health system, not just hospitalized patients.

We assessed these models by performing a reliability audit (model performance and calibration) and fairness audit (summary statistics, subgroup performance, subgroup calibration) to ascertain whether the Epic EOL Index appropriately prioritize patients for ACP in *Primary Care*, and which of the two models appropriately prioritizes patients for ACP in *Inpatient Oncology* and *Hospital Medicine*.

Methods

We first provide details on the two models and then summarize the processes required to complete the fairness and reliability audit. We describe the metrics that comprised the quantitative aspect of the audit. We then describe the methods we used to identify and gather the data needed to complete the audit, including calculating the minimum sample size of ground truth labels required for model evaluation, obtaining those ground truth labels by clinician review, and merging those labels with patient records to create the audit dataset. Lastly, we describe the methods used to compute the audit metrics, and how we presented the results of the audit to clinicians to obtain feedback.

AI models

We audited two models currently deployed at Stanford Health Care: the Epic EOL Index model and Stanford HM ACP model (**Table 1**).

The Epic EOL Index model (38) is a logistic regression model that predicts risk of 12-month mortality (**Table 1**). It takes in 46 input features including demographics (e.g., age, sex, insurance status), labs (e.g., albumin, RDW), comorbidities (e.g., such as those relating to cancer, neurological diagnoses, cardiologic diagnoses, and more), and medications. While organizations using the Epic EHR software are able to set any threshold for converting the model output into a flag to indicate an action is recommended, two thresholds are pre-specified by Epic: a *low threshold* of 0.15 selected based on sensitivity (38), and a *high threshold* of 0.45 selected based on positive predictive value (38). We decided to audit the Epic EOL Index with the low threshold in Primary Care (given lower patient acuity) and with the high threshold in Inpatient Oncology and Hospital Medicine. We retrieved scores on 16 November 2021 for Primary Care, 14 June 2021 for Inpatient Oncology and 31 January 2022 for Hospital Medicine.

The Stanford HM ACP model is a gradient boosted tree model (39) that predicts risk of 3–12 month mortality (**Table 1**). It takes 13,189 input features including demographics (e.g., age, sex), lab orders (e.g., complete blood count with differential, arterial blood gas) and procedure orders (e.g., ventilation, respiratory nebulizer) for all hospitalizations within the last year and is run daily on patients admitted to the Hospital. Patients with a model output probability above 0.25 are flagged in a "Recommended for Advance Care Planning" column in Epic available to all clinicians at Stanford (25, 26). On a retrospective cohort involving 5,965 patients with 12-month mortality labels (prevalence of 24%), this model flagged 23% of patients and had a PPV of 61% (25). For Inpatient Oncology and Hospital Medicine, we retrieved scores for patients on the day of the clinician's label for that patient.

Audit metrics

In previous work (22) we synthesized items that were suggested for reporting by model reporting guidelines to identify the most relevant items for reliability and fairness.

To quantify model reliability, we computed sensitivity, specificity and PPV as these estimate a model's diagnostic capabilities. We computed 95% confidence intervals for each of these metrics using the empirical bootstrap (40) with 1,000 bootstrap samples. We also assessed model calibration using calibration plots and the Observed events/Expected events (O/E) ratio (see details below in the section titled Performing the Audit).

To quantify model fairness, we computed summary statistics across subgroups, defined by sex, race/ethnicity, and age as well as the intersection of race/ethnicity and sex. We also evaluated the model's performance metrics and calibration in each of these subgroups (see details below in the section titled Performing the Audit).

Gathering the data required for the audit

Sample size calculation

We calculated a minimum necessary sample size for external validation of the two prediction models, based on a desired level of calibration (41). We measured calibration as O/E and used the delta method for computing a confidence

TABLE 1 Model information for each setting.

Setting	Primary Care	Inpatient Oncology	Inpatient Oncology	Hospital Medicine	Hospital Medicine
Model	Epic EOL – Low Threshold	Epic EOL – High Threshold	Stanford HM ACP	Epic EOL – High Threshold	Stanford HM ACP
Features	Demographics (Age, Sex, Insurance status), Labs (Albumin, RDW), Comorbidities (Cancer, Neuro., Psych., … Cardio., Resp., …), Medications (many)	Demographics (Age, Sex, Insurance status), Labs (Albumin, RDW), Comorbidities (Cancer, Neuro., Psych., … Cardio., Resp., …), Medications (many)	Demographics (Age, Sex), Lab/Procedure Orders (done in the last year)	Demographics (Age, Sex, Insurance status), Labs (Albumin, RDW), Comorbidities (Cancer, Neuro., Psych., … Cardio., Resp., …), Medications (many)	Demographics (Age, Sex), Lab/Procedure Orders (done in the last year)
# Features	46	46	13,189	46	13189
Model Type	Logistic Regression	Logistic Regression	Gradient boosted Tree	Logistic Regression	Gradient boosted Tree
Output	One-year Mortality Risk	One-year Mortality Risk	One-year Mortality Risk	One-year Mortality Risk	One-year Mortality Risk
Predictions Available For:	All adult patients within health system	All adult patients within health system	All currently hospitalized adult patients	All adult patients within health system	All currently hospitalized adult patients
Threshold	0.15 (Low)	0.45 (High)	0.25 (HM Implementation Threshold)	0.45 (High)	0.25 (HM Implementation Threshold)
Source of Model Information	Epic Cognitive Computing Model Brief: End of Life Index (Galaxy, PDF)	Epic Cognitive Computing Model Brief: End of Life Index (Galaxy, PDF)	AI ACP Technical Details	Epic Cognitive Computing Model Brief: End of Life Index (Galaxy, PDF)	AI ACP Technical Details
Time of Model Predictions	11/16/2021	6/14/2021	8/15/2021–3/19/2022	1/31/2022	2/21/2022, 2/23/2022, 3/1/2022, 3/4/2022, 3/7/2022, 3/14/2022, 3/21/2022
Notes on Time of Model Predictions	Daily predictions are performed, but were not available to be extracted or retrospectively pulled, so we only used a one-time pull on 11/16/2021	Daily predictions are performed, but were not available to be extracted or retrospectively pulled, so we only used a one-time pull on 6/14/2021	Daily predictions were performed and the most recent model prediction on or before the date of the clinician label was used.	Daily predictions are performed, but were not available to be extracted or retrospectively pulled, so we only used a one-time pull on 1/31/2022	Daily predictions were performed and were stored before sending out email requesting clinicians to label.
Location of Model Predictions	Box Folder: Epic EoL Index Validation	Box Folder: Epic EoL Index Validation	shahlab secure server: /data4/AI-ACP/ predictions/ngb_hist	Box Folder: Epic EoL Index Validation	shahlab secure server: /data4/AI-ACP/ predictions/ngb_hist

interval for O/E (41). Assuming a perfect O/E value being 1.0, we aimed for a 95% confidence interval width of [0.74, 1.34]. Based on clinician feedback, in Primary Care, we assumed a 20% prevalence of the positive label; in Inpatient Oncology, we assumed a 70% prevalence of the positive label. In Hospital Medicine, we assumed a 40% prevalence of the positive label.

Obtaining ground truth labels

We used a validated instrument, the *surprise question* (42), to assign ground truth labels for patients. The surprise question asks "Would you be surprised if [patient X] passed away in [Y years]?" An answer of "no" to the surprise question for a given patient constitutes a positive label (for example, if the treating physician would not be surprised if a patient died in 1 year, we assume that the patient is at high risk of dying and should be labeled as "recommended for advance care planning"). A recent meta-analysis (43) found that among 16 studies, the 6-to-12-month surprise question's sensitivity (using records of 12-month mortality as ground truth) ranged from 12% to 93%; specificity ranged from 14% to 98%, PPV ranged from 15% to 79%, and c-statistic ranged from 0.51 to 0.82. In other words, we used the answer to the surprise question as a *proxy* for Y-year mortality in our patient population, because waiting the Y years to ascertain whether patients passed away would have greatly extended the timeframe required to complete the audit. Our audit thus assessed model performance based on concordance of model predictions with clinician-generated assessments of patient mortality *via* the surprise question.

We specified Y = 1 year for the surprise question for Inpatient Oncology and Hospital Medicine patients and Y = 2 years for the Primary Care setting, given lower acuity of patients in Primary Care clinics (**Table 2**).

To obtain answers to the surprise question for Primary Care patients, we first selected from patients who had a visit with a provider between 7 October 2021 and 7 January 2022. We then randomly sampled 5 unique patients to generate a list for each provider; if there were fewer than 5 unique patients, all patients were kept in the provider's list. We then sent personalized messages using our EHR's messaging system to each provider asking them to answer the surprise question for each randomly selected patient (**Table 3**, **Supplementary Figure S1**). For Hospital Medicine, we identified providers who were on service between 21 February 2022 and 21 March

TABLE 2 Clinician label information.

Setting	Primary Care	Inpatient Oncology	Inpatient Oncology	Hospital Medicine	Hospital Medicine
Model	Epic EOL – Low Threshold	Epic EOL – High Threshold	Stanford HM ACP	Epic EOL – High Threshold	Stanford HM ACP
Clinician Label	2-year Surprise Question	1-year Surprise Question	1-year Surprise Question	1-year Surprise Question	1-year Surprise Question
Time of Clinician Labels	2/11/2022–3/7/2022	8/15/2021–3/19/2022	8/15/2021–3/19/2022	2/21/2022–3/22/2022	2/21/2022–3/22/2022
Clinician Population	All Primary Care clinician faculty at Department of Primary Care and Population Health	2 Oncology attending physicians/faculty at Stanford's (ARK, KR)	2 Oncology attending physicians/faculty at Stanford's (ARK, KR)	Every Hospital Medicine attending physician on service during 2/21/2022–3/22/2022	Every Hospital Medicine attending physician on service during 2/21/2022–3/22/2022
Blinding of Clinicians to Model Predictions	Clinicians were blinded to Epic EOL (the model predictions were not available in the EHR). However, clinicians were not specfically blinded from the Stanford HM ACP model (which was available as a flag in Epic).	Clinicians were blinded to Epic EOL (the model predictions were not available in the EHR). However, clinicians were not specfically blinded from the Stanford HM ACP model (which was available as a flag in Epic).	Clinicians were blinded to Epic EOL (the model predictions were not available in the EHR). However, clinicians were not specfically blinded from the Stanford HM ACP model (which was available as a flag in Epic).	Clinicians were blinded to Epic EOL (the model predictions were not available in the EHR). However, clinicians were not specfically blinded from the Stanford HM ACP model (which was available as a flag in Epic).	Clinicians were blinded to Epic EOL (the model predictions were not available in the EHR). However, clinicians were not specfically blinded from the Stanford HM ACP model (which was available as a flag in Epic).
Unit of Data Set	A clinician's Surprise Question Label for a randomly selected patient within the clinician's panel who had a recent visit with the clinician within the last 3 months	A physician's Surprise Question Label for a patient they are responsible for while they are on service	A physician's Surprise Question Label for a patient they are responsible for while they are on service	A physician's Surprise Question Label for a patient they are responsible for on the day of solicitation	A physician's Surprise Question Label for a patient they are responsible for on the day of solicitation

2022, and sent them a message once a week during that period requesting them to answer the surprise question for the patients they had been responsible for during their shifts in that period (**Table 3**, **Supplementary Figure S2**). For both Primary Care and Hospital Medicine, we incentivize providers to answer the surprise question by offering chocolates to those who received the message. For Inpatient Oncology we selected patients who were seen by either co-author ARK or KR between 15 August 2021 and 19 March 2022. ARK and KR answered the 1-year surprise question for all patients they were responsible for while on hospital service during that period (**Table 2**).

Note that the physicians were blinded to Epic EOL Index model predictions, but they were not blinded to the Stanford HM ACP Flag as the flag was available in Epic and in active use at the time of the audit. Co-author ARK reported occasionally referencing the flag when answering the surprise question for patients with rarer cancers. While we recognize this biases our results in favor of the Stanford HM ACP model, we also did not have the ability to suppress the flag just for those clinicians.

Creating the audit data set

Each patient's surprise question ground truth labels were linked with their corresponding patient records from our clinical data warehouse (44), which included patient demographics (sex, date of birth, race, ethnicity), and with the two models' output predictions (**Figure 1**).

We excluded all patients where their provider had not answered the surprise question during the response period. For Inpatient Oncology, we also excluded all patients for which a medical record number was not available. The number of patients excluded for these reasons are provided in the Results.

Finally, we converted patient demographic data into one-hot encoded columns. For sex, we assigned this value based on biological sex (45) (and did one-hot encodings of the potential values). For age, we computed the patient's age at the time of the clinician's surprise question assessment by subtracting their date of birth; we then generated age subgroups by decade of life, e.g., (10, 20], (20, 30], etc. For ethnicity/race, we pulled the ethnicity variable and the race variable, both based on Office of Management and Budget variables (46). We then performed one-hot encoding of the ethnicity and race variables separately, and used a logical AND to generate the ethnicity/race variable: e.g., a Hispanic or Latino, White patient. Lastly, for ethnicity/race and sex, we created intersectional combinations using a logical AND to identify all observed permutations of these variables.

Performing the audit

After we generated the audit data set, we first computed summary statistics. Specifically, for each demographic variable (sex, age, ethnicity/race, and the intersection of ethnicity/race and sex), we computed the counts of each subgroup within

TABLE 3 Solicitation of clinician labels.

Setting	Primary Care	Inpatient Oncology	Inpatient Oncology	Hospital Medicine	Hospital Medicine
Model	Epic EOL – Low Threshold	Epic EOL – High Threshold	Stanford HM ACP	Epic EOL – High Threshold	Stanford HM ACP
Sample Size Required to achieve calibration 95% O/E CI of [0.74, 1.34] (assuming true O/E = 1)	176 assuming prevalence of 20%	19 assuming prevalence of 70%	19 assuming prevalence of 70%	66 assuming prevalence of 40%	66 assuming prevalence of 40%
Solicitation of Clinician Labels	Epic Staff Message sent 2/11/2022	N/A (Physician answered surprise question for all patients responsible for each morning on service)	N/A (Physician answered surprise question for all patients each morning on service)	Secure Emails sent 2/21/2022, 2/23/2022, 3/1/2022, 3/4/2022, 3/7/2022, 3/14/2022, 3/21/2022	Secure Emails sent 2/21/2022, 2/23/2022, 3/1/2022, 3/4/2022, 3/7/2022, 3/14/2022, 3/21/2022
Generation of Solicitations	1. Link visits at a primary care visit site since 09/2021 with patient demographics 2. Filter to visits after 10/72/2021 3. For each provider: filter to visits with the provider that were with patients within their panel 4. Remove visits for providers on days where that provider had more than 30 visits (assume this is artifact of data base) 5. Randomly sample 5 patients of remaining	N/A	N/A	1. For each attending physician on service, generate an email asking them to answer the surprise question for all patients they are responsible for that day	1. For each attending physician on service, generate an email asking them to answer the surprise question for all patients they are responsible for that day
Example Solicitation	Link	N/A	N/A	Link	Link
Announcement of Solicitation	Slide in Division Meeting	N/A	N/A	Email at week start	Email at week start
Incentive with Solicitation	Bag of Ghirardelli Chocolates personally addressed, thanking for answering the surprise question	N/A	N/A	Bag of Ghirardelli Chocolates personally addressed, thanking for answering the surprise question	Bag of Ghirardelli Chocolates personally addressed, thanking for answering the surprise question
Location of Code to Generate Solicitations	shahlab secure server: /data4/jhlu/EOL/[2022-02-01 using concept] pcph_merge_visits_generate_validation_lists_and_plausibility_lists.ipynb	N/A	N/A	shahlab secure server: /data4/jhlu/hm-surprise-gathering/PROD	shahlab secure server: /data4/jhlu/hm-surprise-gathering/PROD
# Clinicians Solicited	79	N/A	N/A	22	22
Size of Solicitations	386	N/A	N/A	545	545

that demographic, as well as the % of the count within the entire data set, and the number and % of positive ground truth labels. We also computed a 95% confidence interval on the positive ground truth label prevalence in each subgroup, using the Clopper-Pearson interval (47) and determined if it overlapped with the confidence interval of the overall positive label prevalence; this evaluated whether ground truth labels were consistent across different demographic subgroups.

We next evaluated model performance. With the ground truth labels and model flags, we computed the following metrics: number of flagged patients, PPV, sensitivity, and specificity. For completeness, we also include the AUROC and Accuracy in the **Supplementary Results**, but do not focus on these in the main text as the other metrics were considered more clinically and diagnostically relevant. We computed 95% confidence intervals on the performance metrics using the empirical bootstrap: we generated 1,000 bootstrap samples of the data set. For each sample, we computed the performance metrics, and computed the difference between each metric from the bootstrap sample and that from the overall study group. (Note the metric on the bootstrap sample may have been null due to dividing by zero, e.g., for PPV if there were no patients that were flagged by the model) We used these differences to generate a distribution of 1,000 bootstrap differences, computed the 2.5th and 97.5th percentiles of the differences (excluding null values), and subtracted these from each metric to generate the empirical bootstrap confidence interval for each metric.

We also evaluated model performance for the subgroups defined by the demographic variables above by computing PPV, sensitivity, and specificity. We computed 95% confidence intervals for each subgroup as above, replacing "overall study group" with the subgroup. We then check if the confidence intervals overlap. Note that resulting confidence intervals had values in some cases that were above 1 or below 0, due to large differences resulting from wide variation in the metric over the bootstrap sampling (40).

We evaluated the models' calibration using calibration plots. A calibration plot provides a visual assessment of how well predicted risk probabilities are aligned with observed outcomes. To generate the calibration plots, we grouped predicted probabilities into quintiles, and within each quintile, computed the average of the predicted risks. We then plotted the averaged predicted risk for each quintile on the x-axis and proportion of positive ground truth labels for each quintile on the y-axis (6, 8–10, 12, 14). We also computed the Observed events/Expected events ratio O/E, which measures the overall calibration of risk predictions, which is computed as the ratio of the total number of observed to predicted events. We computed O/E by dividing the total number of positive ground truth labels by the sum of model output probabilities and used the delta method for computing a 95% confidence interval on O/E (50). The ideal value for O/E is 1; a value <1 or >1 implies that the model over or under predicts the number of events, respectively (41).

We evaluated subgroup calibration by generating calibration plots and by computing the O/E for each subgroup, again using the delta method to compute a 95% confidence interval on O/E (50). Note: because this method's standard error formula for ln(O/E) has O in the denominator, the interval is undefined if O = 0.

Presenting audit results to decision makers

We presented the results of our audit to decision makers in Primary Care (co-authors AS, WT), Inpatient Oncology (co-authors ARK, WT, SC, KR, MG), and Hospital Medicine (co-authors SW, LS, RL), in a separate presentation for each setting. Each presentation first gave context to the audit, including sharing previous findings that AI models have been unreliable (5, 48) or unfair (4), as well as that race/ethnicity data in the EHR is known to have inaccuracies (33). Then, we shared the summary statistics, model performance, model calibration, subgroup performance and subgroup calibration.

We also designed a survey for the decision makers to complete at the end of each presentation (**Supplementary Methods**). In the survey, we assessed their understanding of reliability/fairness by asking "What does it mean to you for a model to be reliable/fair?" and "What are the first thoughts that came to your mind on seeing the results of the reliability and fairness audit?" We also assessed whether specific components of the reliability/fairness audit would or would not affect decision making, and asked if there would be any other information they believe should be included in the audit. Example surveys were shared with several decision makers (co-authors WT, SW, AS), informaticists (co-authors AG, AC) and the director of operations of an AI research & implementation team (co-author MS) for feedback prior to giving the survey.

After we received the survey responses, we reviewed and summarized the most common structured responses. We also read the free text responses, identified themes (ensuring that every response had at least one theme represented) and categorized responses by the themes. JL was the sole coder, and performed inductive thematic analysis to generate codes.

Results

Reliability and fairness audit

We report the reliability and fairness audits below. For simplicity, all confidence intervals are listed in the tables. Also, only statistically significant results are listed in the tables; full results including those without statistically significant differences are listed in the **Supplementary Tables**.

Primary Care

We calculated we would need a sample size of 176 to achieve an O/E 95% confidence interval of [0.74, 1.34], assuming a 20% prevalence of the positive label. We solicited 79 clinicians for 386 labels of their patients (2-year surprise question answers). 70 clinicians responded with 344 labels (89% response rate). Six of the response labels were "Y/N" or "DECEASED" and were filtered out, leaving 338 labels fitting the schema.

Epic EOL Low Threshold in Primary Care

The final data set size for the Epic EOL – Low Threshold model in Primary Care was 338 with 68 positive labels after we linked the 338 clinician labels fitting the schema with Epic EOL model predictions and patient demographics (**Table 4**).

The overall prevalence was 0.2. There was significantly higher prevalence for Age: (80, 90] at 0.55. There was

TABLE 4 Processing and final data sets.

Setting	Primary Care	Inpatient Oncology	Inpatient Oncology	Hospital Medicine	Hospital Medicine
Model	Epic EOL – Low Threshold	Epic EOL – High Threshold	Stanford HM ACP	Epic EOL – High Threshold	Stanford HM ACP
Location of Gathered Clinician Labels	Box file	Box file	Box file	Box folder	Box folder
# Clinicians Responding	70	2	2	18	18
Size of Clinician Labels (raw)	344	225	225	413	413
Clinician Labels/Solicitations (%)	89%	N/A	N/A	76%	76%
Missing Clinician Labels	42	N/A	N/A	132	132
Size of Clinician Labels Fitting Schema	338	202	202	409	409
# Outcomes in Clinician Labels Fitting Schema	68	136	136	178	178
% Outcomes in Clinician Labels Fitting Schema	20%	67%	67%	44%	44%
Clinician Labels not fitting schema	4 – "Y/N" 2 – "DECEASED"	23 – Not linked to numerical MRN	23 – Not linked to numerical MRN	2 – "TRANSFERRED" 2 – "Maybe"	2 – "TRANSFERRED" 2 – "Maybe"
Final Data Set Size (has Clinician Label, Model Prediction, and Demographics)	338	150	115	305	225
# Outcomes in Final Data Set	68	105	79	133	99
% Outcomes in Final Data Set	20%	70%	69%	44%	44%

significantly lower prevalence for Age: (20, 30] at 0 and Age: (30, 40] at 0. There were no significant differences in prevalence found by Sex, Ethnicity/Race, or the intersection of Ethnicity/Race and Sex (Table 5, Supplementary Tables S1–S4).

The model flagged 30 patients out of 338 (9%), exhibiting low sensitivity (0.37), high specificity (0.98), and high PPV (0.83). The model also underpredicted events relative to clinicians by a factor of O/E = 4.1. There was significantly lower sensitivity for Age: (60, 70] at 0.1 and Age: (70, 80] at 0.07. The model also underpredicted events more for Age: (60, 70], by a factor of O/E = 9.3 (Table 5). For several other groups, there were statistically significant differences in prevalence, performance or O/E, but these subgroups had less than 10 patients to calculate the metric for, making results inconclusive (Table 5).

Inpatient Oncology

We calculated we would need a sample size of 19 to achieve an O/E 95% confidence interval of [0.74, 1.34], assuming a 70% prevalence of the positive label. Two clinicians (ARK, KR) completed 225 labels for patients they saw while on service (1-year surprise question answers). Note: each data point corresponds with a unique patient encounter (some patients were included multiple times due to re-hospitalization). Of the 225 labels, 23 did not have a numerical MRN associated and were filtered out, leaving 202 clinician labels fitting the schema.

Epic EOL High Threshold in Inpatient Oncology

The final data set size for the Epic EOL – High Threshold model in Inpatient Oncology, was 150 with 105 positive labels after we linked the 202 clinician labels fitting the schema with Epic EOL model predictions and patient demographics (Table 4).

The overall prevalence was 0.7. There was significantly lower prevalence for younger patients (0.23 for Age: (20, 30]). There were no significant differences in prevalence by Sex, Ethnicity/Race, and the intersection of Ethnicity/Race and Sex (Table 6).

The model flagged 32 patients out of 150 (21%) with a sensitivity of 0.27, specificity of 0.91, and PPV of 0.88. The model predicted many fewer events relative to the number of positive clinician labels, with an O/E ratio of 3. Sensitivity for Hispanic or Latino patients with Race "Other" (0.09) was significantly lower than the model's overall sensitivity (0.27). This was also true for Hispanic or Latino Males with Race "Other" specifically, for which the model's sensitivity was 0. The model significantly underpredicted events for both subgroups relative to clinicians, with O/E ratios of 6.9 and 9, respectively. Several other subgroups exhibited statistically significant differences in model performance or O/E, but these subgroups had less than 10 patients to calculate the metric for, making such claims inconclusive. See Table 6 for details.

Stanford HM ACP in Inpatient Oncology

The final data set size for the Stanford HM ACP model in Inpatient Oncology was 114 with 79 positive labels after we linked the 202 clinician labels fitting the schema with

TABLE 5 Epic EOL Low threshold in primary care: reliability and fairness audit with significant results. Prevalence, performance and calibration is presented for the overall cohort and for subgroups with significant differences in prevalence, significantly lower performance, or significantly higher O/E (bolded). For the full set of results, see **Supplementary Tables S1–S4**.

Group	Sample Size	Prevalence (Fraction)	Prevalence [95% CI]	Sensitivity (Fraction)	Sensitivity [95% CI]	Specificity (Fraction)	Specificity [95% CI]	Positive Predictive Value (Fraction)	Positive Predictive Value [95% CI]	O/E (Fraction)	O/E [95% CI]
Overall	338	0.2 (68/338)	[0.16, 0.25]	0.37 (25/68)	[0.26, 0.49]	0.98 (265/270)	[0.97, 1.0]	0.83 (25/30)	[0.7, 0.98]	4.1 (68/16.4)	[3.3, 5.1]
Age: (20, 30]	27	**0.0 (0/27)**	**[0, 0.13]**	nan (0/0)	N/A	1.0 (1/1)	[1.0, 1.0]	nan (0/0)	N/A	nan (0/0.0)	N/A
Age: (30, 40]	61	**0.0 (0/61)**	**[0, 0.06]**	nan (0/0)	N/A	1.0 (1/1)	[1.0, 1.0]	nan (0/0)	N/A	0.0 (0/0.0)	N/A
Age: (50, 60]	48	0.04 (2/48)	[0.01, 0.14]	**0.0 (0/2)**	**[0.0, 0.0]**	1.0 (46/46)	[1.0, 1.0]	nan (0/0)	N/A	4.7 (2/0.4)	[1.2, 18.1]
Age: (60, 70]	51	0.2 (10/51)	[0.1, 0.33]	**0.1 (1/10)**	**[-0.13, 0.2]**	1.0 (41/41)	[1.0, 1.0]	1.0 (1/1)	[1.0, 1.0]	**9.3 (10/1.1)**	**[5.3, 16.1]**
Age: (70, 80]	51	0.29 (15/51)	[0.17, 0.44]	**0.07 (1/15)**	**[-0.09, 0.13]**	0.97 (35/36)	[0.94, 1.03]	0.5 (1/2)	[0.0, 1.0]	6.3 (15/2.4)	[4.1, 9.6]
Age: (80, 90]	33	**0.55 (18/33)**	**[0.36, 0.72]**	0.39 (7/18)	[0.15, 0.61]	0.87 (13/15)	[0.73, 1.07]	0.78 (7/9)	[0.56, 1.11]	3.4 (18/5.3)	[2.5, 4.6]
Age: (90, 100]	19	**0.84 (16/19)**	**[0.6, 0.97]**	0.81 (13/16)	[0.62, 1.0]	**0.33 (1/3)**	**[-0.33, 0.67]**	0.87 (13/15)	[0.73, 1.05]	2.7 (16/5.9)	[2.2, 3.3]
Ethnicity: Hispanic or Latino, Race: Other	20	0.05 (1/20)	[0.0, 0.25]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (19/19)	[1.0, 1.0]	nan (0/0)	N/A	4.2 (1/0.2)	[0.6, 28.1]
Ethnicity: Not Hispanic or Latino, Race: Native Hawaiian or Other Pacific Islander	3	0.33 (1/3)	[0.01, 0.91]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (2/2)	[1.0, 1.0]	nan (0/0)	N/A	**50.0 (1/0.0)**	**[10.1, 247.7]**
Ethnicity: Hispanic or Latino, Race: Other, Sex: Male	9	0.11 (1/9)	[0.0, 0.48]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (8/8)	[1.0, 1.0]	nan (0/0)	N/A	7.1 (1/0.1)	[1.1, 45.3]
Ethnicity: Not Hispanic or Latino, Race: Native Hawaiian or Other Pacific Islander, Sex: Female	3	0.33 (1/3)	[0.01, 0.91]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (2/2)	[1.0, 1.0]	nan (0/0)	N/A	**50.0 (1/0.0)**	**[10.1, 247.7]**

TABLE 6 Epic EOL high threshold in inpatient oncology: reliability and fairness audit, significant results. Prevalence, performance and calibration is presented for the overall cohort and for subgroups with significant differences in prevalence, significantly lower performance, or significantly higher O/E (bolded). For the full set of results, see **Supplementary Tables S5–S8**.

Group	Sample Size	Prevalence (Fraction)	Prevalence [95% CI]	Sensitivity (Fraction)	Sensitivity [95% CI]	Specificity (Fraction)	Specificity [95% CI]	Positive Predictive Value (Fraction)	Positive Predictive Value [95% CI]	O/E (Fraction)	O/E [95% CI]
Overall	150	0.7 (105/150)	[0.62, 0.77]	0.27 (28/105)	[0.18, 0.34]	0.91 (41/45)	[0.84, 1.0]	0.88 (28/32)	[0.78, 1.01]	3.0 (105/34.8)	[2.7, 3.4]
Age: (20, 30]	13	**0.23 (3/13)**	**[0.05, 0.54]**	**0.0 (0/3)**	**[0.0, 0.0]**	1.0 (10/10)	[1.0, 1.0]	nan (0/0)	N/A	5.0 (3/0.6)	[1.9, 13.5]
Age: (30, 40]	14	0.57 (8/14)	[0.29, 0.82]	**0.0 (0/8)**	**[0.0, 0.0]**	1.0 (6/6)	[1.0, 1.0]	nan (0/0)	N/A	**7.5 (8/1.1)**	**[4.8, 11.9]**
Age: (60, 70]	34	0.85 (29/34)	[0.69, 0.95]	0.24 (7/29)	[0.07, 0.39]	**0.4 (2/5)**	**[-0.2, 0.8]**	0.7 (7/10)	[0.4, 1.02]	3.0 (29/9.8)	[2.6, 3.4]
Ethnicity: Hispanic or Latino, Race: Other	30	0.73 (22/30)	[0.54, 0.88]	**0.09 (2/22)**	**[-0.05, 0.18]**	1.0 (8/8)	[1.0, 1.0]	1.0 (2/2)	[1.0, 1.0]	**6.9 (22/3.2)**	**[5.6, 8.6]**
Ethnicity: Hispanic or Latino, Race: White	3	0.33 (1/3)	[0.01, 0.91]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (2/2)	[1.0, 1.0]	nan (0/0)	N/A	2.9 (1/0.3)	[0.6, 14.6]
Ethnicity: Hispanic or Latino, Race: Other, Sex: Male	17	0.76 (13/17)	[0.5, 0.93]	**0.0 (0/13)**	**[0.0, 0.0]**	1.0 (4/4)	[1.0, 1.0]	nan (0/0)	N/A	**9.0 (13/1.4)**	**[6.9, 11.8]**
Ethnicity: Hispanic or Latino, Race: Other, Sex: Female	13	0.69 (9/13)	[0.39, 0.91]	0.22 (2/9)	[-0.06, 0.44]	1.0 (4/4)	[1.0, 1.0]	1.0 (2/2)	[1.0, 1.0]	**5.2 (9/1.7)**	**[3.6, 7.4]**
Ethnicity: Not Hispanic or Latino, Race: Other, Sex: Female	5	1.0 (5/5)	[0.48, 1]	0.2 (1/5)	[-0.2, 0.4]	nan (0/0)	N/A	1.0 (1/1)	[1.0, 1.0]	**4.9 (5/1.0)**	**[4.9, 4.9]**
Ethnicity: Not Hispanic or Latino, Race: Black or African American, Sex: Female	2	0.5 (1/2)	[0.01, 0.99]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (1/1)	[1.0, 1.0]	nan (0/0)	N/A	4.2 (1/0.2)	[1.0, 16.7]
Ethnicity: Hispanic or Latino, Race: White, Sex: Female	1	1.0 (1/1)	[0.03, 1]	**0.0 (0/1)**	**[0.0, 0.0]**	nan (0/0)	N/A	nan (0/0)	N/A	**inf (1/0.0)**	**[inf, inf]**

Stanford HM ACP model predictions and patient demographics (Table 4).

The overall prevalence was 0.69. There were no significant differences in prevalence amongst the demographic subgroups considered.

The Stanford HM ACP model flagged 85 patients out of 114 (75%) with sensitivity 0.89, specificity 0.57, and PPV 0.82. The model moderately underestimated events relative to clinicians, with an O/E of 1.7. Model performance and O/E appeared to differ for some subgroups, but these subgroups had less than 10 patients to calculate the metric for, making any associated claims inconclusive. See Table 7 for details.

Model comparison in Inpatient Oncology

Comparing model performance in Inpatient Oncology, the Stanford HM ACP model flagged more patients (75% vs. 21%), had significantly higher sensitivity (0.89 vs. 0.27), and exhibited similar PPV (0.82 vs. 0.88, 95% confidence intervals overlap). The Epic EOL High Threshold model had significantly higher specificity (0.91 vs. 0.57). Comparing model calibration, the Stanford HM ACP model had significantly better calibration in terms of O/E (1.7 vs. 3).

Hospital Medicine

We calculated we would need a sample size of 66 to achieve an O/E confidence interval of [0.74, 1.34], assuming a 40% prevalence of the positive label. We solicited 22 clinicians for 545 labels of their patients seen while they were on service (1-year surprise question answers). 18 clinicians responded with 413 labels (76% response rate). Note: each data point corresponds with a unique patient encounter (some patients were included multiple times due to long hospital stays). Four of these were "Maybe" or "TRANSFERRED" and were filtered out, leaving 409 clinician labels fitting the schema.

Epic EOL High Threshold in Hospital Medicine

The final data set size for the Epic EOL – High Threshold model in Hospital Medicine, was 305 with 133 positive labels after we linked the 409 clinician labels fitting the schema with Epic EOL model predictions and patient demographics (Table 4).

The overall prevalence was 0.44. Prevalence did not differ by sex, but was significantly higher for older patients (0.76 for Age: (80, 90] and 0.94 for Age: (90, 100]) and significantly lower for younger patients (0.12 for Age: (20, 30] and 0.15 for Age: (30, 40]). Prevalence was also significantly higher for Non-Hispanic Asian patients (0.68) but significantly lower for Hispanic or Latino patients with Race "Other" (0.18) and, in particular, Hispanic or Latino Males of Race "Other" (0.14).

The model flagged 34 out of 305 patients (11%). The model demonstrated a sensitivity of 0.2, specificity of 0.95, and PPV of 0.76. The model underpredicted events relative to clinicians (O/E ratio of 2.5). There was significantly lower sensitivity for Age: (50,60] at 0. The model significantly underestimated events relative to clinicians for Non-Hispanic White Females (O/E = 3.7). Differences in performance and O/E were statistically significant for other subgroups, but these subgroups had less than 10 patients to calculate the metric for, preventing conclusive statements regarding disparate performance. See Table 8 for details.

Stanford HM ACP in Hospital Medicine

The final data set size for the Stanford HM ACP model in Hospital Medicine, was 225 with 99 positive labels after we linked the 409 clinician labels fitting the schema with Stanford HM ACP model predictions and patient demographics (Table 4).

The overall prevalence was 0.44. Prevalence was significantly higher for older patients (0.8 for Age: (80, 90], 0.92 for Age: (90, 100]) and significantly lower for younger patients (0.11 for Age: (30, 40]). Prevalence was also significantly lower for Hispanic or Latino patients with Race "Other" (0.16) and significantly higher for Non-Hispanic Asian patients (0.7), especially Non-Hispanic Asian Males (0.81).

The Stanford HM ACP model flagged 85 out of 225 patients (38%), with sensitivity 0.69, specificity 0.87, and PPV 0.8. Relative to clinicians, the model underestimated events by a factor of O/E = 1.5. For patients Age: (90, 100], this underestimation was even more substantial with an O/E ratio of 2.5. Specificity was lower (0.57) for Age: (70, 80]. Relative to the model's overall PPV, the PPV for Hispanic or Latino patients with Race "Other" was significantly lower (0.29 vs. 0.8). Model performance disparities in other subgroups were inconclusive given they had less than 10 patients to calculate the metric for. See Table 9 for details.

Model comparison in Hospital Medicine

Comparing model performance in Hospital Medicine, relative to the Epic EOL – High Threshold model the Stanford HM ACP model flagged more patients (38% vs. 11%), had significantly higher sensitivity (0.69 vs. 0.2), similar specificity (0.87 vs. 0.95, 95% confidence intervals overlap), and similar PPV (0.8 vs. 0.76, 95% confidence intervals overlap). Comparing model calibration, the Stanford HM ACP model had significantly better calibration in O/E (1.5 vs. 2.5).

Supplemental analysis with class balancing

We also performed a supplemental analysis of the reliability/fairness audits after using random oversampling to achieve class balance (see Supplementary Results). Overall, model sensitivity and specificity stayed the same for all settings. Model PPV increased when class balancing increased the prevalence (Primary Care, Hospital Medicine), and decreased when class balancing decreased the prevalence (Inpatient Oncology). Model calibration in O/E had inconsistent changes after class balancing. The differences in performance and calibration between the Epic EOL High

TABLE 7 Stanford HM ACP in inpatient oncology: reliability and fairness audit with significant results. Prevalence, performance and calibration is presented for the overall cohort and for subgroups with significant differences in prevalence, significantly lower performance, or significantly higher O/E (bolded). For the full set of results, see **Supplementary Tables S9–S12**.

Group	Sample Size	Prevalence (Fraction)	Prevalence [95% CI]	Sensitivity (Fraction)	Sensitivity [95% CI]	Specificity (Fraction)	Specificity [95% CI]	Positive Predictive Value (Fraction)	Positive Predictive Value [95% CI]	O/E (Fraction)	O/E [95% CI]
Overall	114	0.69 (79/114)	[0.6, 0.78]	0.89 (70/79)	[0.82, 0.96]	0.57 (20/35)	[0.4, 0.74]	0.82 (70/85)	[0.74, 0.91]	1.7 (79/46.2)	[1.5, 1.9]
Age: (40, 50]	11	0.55 (6/11)	[0.23, 0.83]	0.83 (5/6)	[0.67, 1.17]	**0.2 (1/5)**	**[-0.2, 0.4]**	0.56 (5/9)	[0.24, 0.89]	1.5 (6/4.0)	[0.9, 2.6]
Age: (80, 90]	12	0.83 (10/12)	[0.52, 0.98]	0.9 (9/10)	[0.8, 1.1]	**0.0 (0/2)**	**[0.0, 0.0]**	0.82 (9/11)	[0.64, 1.05]	1.6 (10/6.2)	[1.3, 2.1]
Ethnicity: Hispanic or Latino, Race: White	3	0.33 (1/3)	[0.01, 0.91]	1.0 (1/1)	[1.0, 1.0]	**0.0 (0/2)**	**[0.0, 0.0]**	**0.33 (1/3)**	**[-0.33, 0.67]**	0.8 (1/1.3)	[0.2, 3.9]
Ethnicity: Not Hispanic or Latino, Race: Black or African American	3	0.33 (1/3)	[0.01, 0.91]	**0.0 (0/1)**	**[0.0, 0.0]**	0.5 (1/2)	[0.0, 1.0]	**0.0 (0/1)**	**[0.0, 0.0]**	1.8 (1/0.6)	[0.4, 8.8]
Ethnicity: Not Hispanic or Latino, Race: American Indian or Alaska Native	1	1.0 (1/1)	[0.03, 1]	**0.0 (0/1)**	**[0.0, 0.0]**	nan (0/0)	N/A	nan (0/0)	N/A	**4.1 (1/0.2)**	**[4.1, 4.1]**
Ethnicity: Not Hispanic or Latino, Race: Other, Sex: Female	5	1.0 (5/5)	[0.48, 1]	1.0 (1/1)	[1.0, 1.0]	nan (0/0)	N/A	1.0 (1/1)	[1.0, 1.0]	**2.0 (5/2.4)**	**[2.0, 2.0]**
Ethnicity: Hispanic or Latino, Race: White, Sex: Male	2	0.0 (0/2)	[0, 0.84]	nan (0/0)	N/A	**0.0 (0/2)**	**[0.0, 0.0]**	**0.0 (0/2)**	**[0.0, 0.0]**	0.0 (0/0.5)	N/A
Ethnicity: Not Hispanic or Latino, Race: Black or African American, Sex: Female	2	0.5 (1/2)	[0.01, 0.99]	**0.0 (0/1)**	**[0.0, 0.0]**	**0.0 (0/1)**	**[0.0, 0.0]**	**0.0 (0/1)**	**[0.0, 0.0]**	2.6 (1/0.4)	[0.7, 10.6]
Ethnicity: Not Hispanic or Latino, Race: American Indian or Alaska Native, Sex: Male	1	1.0 (1/1)	[0.03, 1]	**0.0 (0/1)**	**[0.0, 0.0]**	nan (0/0)	N/A	nan (0/0)	N/A	**4.1 (1/0.2)**	**[4.1, 4.1]**

TABLE 8 Epic EOL high threshold in hospital medicine: reliability and fairness audit with significant results. Prevalence, performance and calibration is presented for the overall cohort and for subgroups with significant differences in prevalence, significantly lower performance, or significantly higher O/E (bolded). For the full set of results, see **Supplementary Tables S13–S16**.

Group	Sample Size	Prevalence (Fraction)	Prevalence [95% CI]	Sensitivity (Fraction)	Sensitivity [95% CI]	Specificity (Fraction)	Specificity [95% CI]	Positive Predictive Value (Fraction)	Positive Predictive Value [95% CI]	O/E (Fraction)	O/E [95% CI]
Overall	305	0.44 (133/305)	[0.38, 0.49]	0.2 (26/133)	[0.12, 0.26]	0.95 (164/172)	[0.92, 0.99]	0.76 (26/34)	[0.63, 0.91]	2.5 (133/53.2)	[2.2, 2.8]
Age: (10, 20]	3	0.33 (1/3)	[0.01, 0.91]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (2/2)	[1.0, 1.0]	nan (0/0)	N/A	**inf (1/0.0)**	**[inf, inf]**
Age: (20, 30]	24	**0.12 (3/24)**	**[0.03, 0.32]**	**0.0 (0/3)**	**[0.0, 0.0]**	1.0 (21/21)	[1.0, 1.0]	nan (0/0)	N/A	4.5 (3/0.7)	[1.6, 12.9]
Age: (30, 40]	40	**0.15 (6/40)**	**[0.06, 0.3]**	**0.0 (0/6)**	**[0.0, 0.0]**	1.0 (34/34)	[1.0, 1.0]	nan (0/0)	N/A	4.7 (6/1.3)	[2.2, 9.7]
Age: (50, 60]	40	0.28 (11/40)	[0.15, 0.44]	**0.0 (0/11)**	**[0.0, 0.0]**	1.0 (29/29)	[1.0, 1.0]	nan (0/0)	N/A	3.6 (11/3.1)	[2.2, 5.9]
Age: (80, 90]	34	**0.76 (26/34)**	**[0.59, 0.89]**	0.19 (5/26)	[0.02, 0.34]	0.88 (7/8)	[0.75, 1.18]	0.83 (5/6)	[0.67, 1.17]	2.6 (26/10.0)	[2.2, 3.1]
Age: (90, 100]	18	**0.94 (17/18)**	**[0.73, 1.0]**	0.24 (4/17)	[0.03, 0.41]	**0.0 (0/1)**	**[0.0, 0.0]**	0.8 (4/5)	[0.6, 1.27]	2.2 (17/7.6)	[2.0, 2.5]
Ethnicity: Hispanic or Latino, Race: Other	44	**0.18 (8/44)**	**[0.08, 0.33]**	0.12 (1/8)	[-0.17, 0.25]	0.94 (34/36)	[0.89, 1.02]	0.33 (1/3)	[-0.33, 0.67]	2.0 (8/4.0)	[1.1, 3.7]
Ethnicity: Not Hispanic or Latino, Race: Asian	37	**0.68 (25/37)**	**[0.5, 0.82]**	0.32 (8/25)	[0.12, 0.52]	1.0 (12/12)	[1.0, 1.0]	1.0 (8/8)	[1.0, 1.0]	2.1 (25/12.2)	[1.6, 2.6]
Ethnicity: Hispanic or Latino, Race: White	13	0.23 (3/13)	[0.05, 0.54]	**0.0 (0/3)**	**[0.0, 0.0]**	1.0 (10/10)	[1.0, 1.0]	nan (0/0)	N/A	4.4 (3/0.7)	[1.6, 11.9]
Ethnicity: Not Hispanic or Latino, Race: Native Hawaiian or Other Pacific Islander	10	0.4 (4/10)	[0.12, 0.74]	**0.0 (0/4)**	**[0.0, 0.0]**	1.0 (6/6)	[1.0, 1.0]	nan (0/0)	N/A	3.0 (4/1.3)	[1.4, 6.4]
Ethnicity: Not Hispanic or Latino, Race: White, Sex: Female	64	0.52 (33/64)	[0.39, 0.64]	0.12 (4/33)	[0.01, 0.22]	0.97 (30/31)	[0.94, 1.04]	0.8 (4/5)	[0.6, 1.27]	**3.7 (33/9.0)**	**[2.9, 4.6]**
Ethnicity: Hispanic or Latino, Race: Other, Sex: Male	22	**0.14 (3/22)**	**[0.03, 0.35]**	**0.0 (0/3)**	**[0.0, 0.0]**	1.0 (19/19)	[1.0, 1.0]	nan (0/0)	N/A	4.3 (3/0.7)	[1.5, 12.4]
Ethnicity: Not Hispanic or Latino, Race: Other, Sex: Male	9	0.56 (5/9)	[0.21, 0.86]	**0.0 (0/5)**	**[0.0, 0.0]**	1.0 (4/4)	[1.0, 1.0]	nan (0/0)	N/A	**13.9 (5/0.4)**	**[7.7, 24.9]**
Ethnicity: Hispanic or Latino, Race: White, Sex: Male	7	0.14 (1/7)	[0.0, 0.58]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (6/6)	[1.0, 1.0]	nan (0/0)	N/A	10.0 (1/0.1)	[1.6, 61.4]
Ethnicity: Hispanic or Latino, Race: White, Sex: Female	6	0.33 (2/6)	[0.04, 0.78]	**0.0 (0/2)**	**[0.0, 0.0]**	1.0 (4/4)	[1.0, 1.0]	nan (0/0)	N/A	3.4 (2/0.6)	[1.1, 10.7]

(*continued*)

TABLE 8 Continued

Group	Sample Size	Prevalence (Fraction)	Prevalence [95% CI]	Sensitivity (Fraction)	Sensitivity [95% CI]	Specificity (Fraction)	Specificity [95% CI]	Positive Predictive Value (Fraction)	Positive Predictive Value [95% CI]	O/E (Fraction)	O/E [95% CI]
Ethnicity: Not Hispanic or Latino, Race: Native Hawaiian or Other Pacific Islander, Sex: Female	6	0.17 (1/6)	[0.0, 0.64]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (5/5)	[1.0, 1.0]	nan (0/0)	N/A	7.7 (1/0.1)	[1.3, 46.0]
Ethnicity: Not Hispanic or Latino, Race: Native Hawaiian or Other Pacific Islander, Sex: Male	4	0.75 (3/4)	[0.19, 0.99]	**0.0 (0/3)**	**[0.0, 0.0]**	1.0 (1/1)	[1.0, 1.0]	nan (0/0)	N/A	2.5 (3/1.2)	[1.4, 4.4]

TABLE 9 Stanford HM ACP in hospital medicine: reliability and fairness audit with significant results. Prevalence, performance and calibration is presented for the overall cohort and for subgroups with significant differences in prevalence, significantly lower performance, or significantly higher O/E (bolded). For the full set of results, see **Supplementary Tables S17–S20**.

Group	Sample Size	Prevalence (Fraction)	Prevalence [95% CI]	Sensitivity (Fraction)	Sensitivity [95% CI]	Specificity (Fraction)	Specificity [95% CI]	Positive Predictive Value (Fraction)	Positive Predictive Value [95% CI]	O/E (Fraction)	O/E [95% CI]
Overall	225	0.44 (99/225)	[0.37, 0.51]	0.69 (68/99)	[0.6, 0.78]	0.87 (109/126)	[0.81, 0.93]	0.8 (68/85)	[0.72, 0.89]	1.5 (99/65.2)	[1.3, 1.8]
Age: (10, 20]	3	0.33 (1/3)	[0.01, 0.91]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (2/2)	[1.0, 1.0]	nan (0/0)	N/A	2.4 (1/0.4)	[0.5, 12.1]
Age: (30, 40]	28	**0.11 (3/28)**	**[0.02, 0.28]**	**0.0 (0/3)**	**[0.0, 0.0]**	1.0 (25/25)	[1.0, 1.0]	nan (0/0)	N/A	0.8 (3/3.7)	[0.3, 2.3]
Age: (50, 60]	25	0.24 (6/25)	[0.09, 0.45]	**0.17 (1/6)**	**[-0.17, 0.33]**	1.0 (19/19)	[1.0, 1.0]	1.0 (1/1)	[1.0, 1.0]	1.5 (6/4.1)	[0.7, 2.9]
Age: (70, 80]	48	0.56 (27/48)	[0.41, 0.71]	0.81 (22/27)	[0.67, 0.99]	**0.57 (12/21)**	**[0.37, 0.78]**	0.71 (22/31)	[0.56, 0.88]	1.3 (27/20.2)	[1.0, 1.7]
Age: (80, 90]	30	**0.8 (24/30)**	**[0.61, 0.92]**	0.75 (18/24)	[0.59, 0.93]	0.5 (3/6)	[0.0, 1.0]	0.86 (18/21)	[0.71, 1.01]	2.1 (24/11.7)	[1.7, 2.5]
Age: (90, 100]	13	**0.92 (12/13)**	**[0.64, 1.0]**	0.83 (10/12)	[0.67, 1.05]	1.0 (1/1)	[1.0, 1.0]	1.0 (10/10)	[1.0, 1.0]	**2.5 (12/4.9)**	**[2.1, 2.9]**
Ethnicity: Hispanic or Latino, Race: Other	38	**0.16 (6/38)**	**[0.06, 0.31]**	0.33 (2/6)	[-0.13, 0.67]	0.84 (27/32)	[0.72, 0.96]	**0.29 (2/7)**	**[-0.1, 0.57]**	0.9 (6/7.0)	[0.4, 1.8]
Ethnicity: Not Hispanic or Latino, Race: Asian	37	**0.7 (26/37)**	**[0.53, 0.84]**	0.73 (19/26)	[0.58, 0.91]	0.91 (10/11)	[0.82, 1.13]	0.95 (19/20)	[0.9, 1.07]	1.6 (26/16.1)	[1.3, 2.0]
Ethnicity: Not Hispanic or Latino, Race: Asian, Sex: Male	21	**0.81 (17/21)**	**[0.58, 0.95]**	0.65 (11/17)	[0.42, 0.87]	0.75 (3/4)	[0.5, 1.25]	0.92 (11/12)	[0.83, 1.12]	1.8 (17/9.6)	[1.4, 2.2]
Ethnicity: Hispanic or Latino, Race: Other, Sex: Male	16	0.12 (2/16)	[0.02, 0.38]	**0.0 (0/2)**	**[0.0, 0.0]**	0.86 (12/14)	[0.71, 1.05]	**0.0 (0/2)**	**[0.0, 0.0]**	0.9 (2/2.2)	[0.3, 3.4]
Ethnicity: Not Hispanic or Latino, Race: Black or African American, Sex: Male	9	0.11 (1/9)	[0.0, 0.48]	**0.0 (0/1)**	**[0.0, 0.0]**	0.88 (7/8)	[0.75, 1.12]	**0.0 (0/1)**	**[0.0, 0.0]**	0.8 (1/1.3)	[0.1, 5.0]
Ethnicity: Not Hispanic or Latino, Race: Native Hawaiian or Other Pacific Islander, Sex: Female	6	0.17 (1/6)	[0.0, 0.64]	**0.0 (0/1)**	**[0.0, 0.0]**	1.0 (5/5)	[1.0, 1.0]	nan (0/0)	N/A	0.8 (1/1.2)	[0.1, 5.0]
Ethnicity: Not Hispanic or Latino, Race: Other, Sex: Male	6	0.5 (3/6)	[0.12, 0.88]	**0.0 (0/3)**	**[0.0, 0.0]**	1.0 (3/3)	[1.0, 1.0]	nan (0/0)	N/A	2.3 (3/1.3)	[1.0, 5.0]

Threshold model and the Stanford HM ACP model stayed the same in each setting in the class balance analysis. Some of the subgroup differences in prevalence, performance and calibration were maintained in the class balance analysis. Overall, interpretation of the results after class balancing is difficult given that class balancing can lead to poorly calibrated models (49, 50).

Survey of decision makers

After the presentations, we administered a survey about how the audit impacted decision makers' decision to use the model. We gathered 10 responses: 2 for Primary Care, 5 for Inpatient Oncology and 3 for Hospital Medicine. 7 responses were from Attending Physicians, 1 was from a Physician Assistant, and 2 were from the Lead for the Serious Illness Care Program.

Understandings of reliable/fair models

Decision makers used themes of **Accurate** (9/10) and **Consistent** (5/10) when asked to describe what it meant to them for a model to be reliable (**Table 10**). For example, one response said: "*not brittle (doesn't give really weird answers if some data are missing).*"

When asked to describe what it meant to them for a model to be fair, they tended to use themes of **Similar Model Performance across demographics** (6/10) often specifically citing **Race/Ethnicity** (4/10) and **Sex** (4/10) (**Table 11**). Another common theme was **Depends on How Model is Used** (2/10). For example, one response said: "*… In one context, being more sensitive for patients of a certain group could be good (fair) for those patients, in another context it could be bad (unfair).*"

Decision makers used a variety of themes to describe their first thoughts on seeing the results of the reliability and fairness audit (**Supplementary Table S21**). In Primary care, the decision makers used **Excitement** and **Trust to Use the Model For Intended Purpose** (2/2), whereas in Hospital Medicine, they used **Interesting** (3/3). In Inpatient Oncology, 2 of 5 responses referred to **Low Sample Size**, for example "*… There may be some signals of differences based on age and race/ethnicity groups, but I wonder if this is in part limited by low power.*"

TABLE 10 Survey responses to "what does it mean to you for a model to be reliable?".

Theme	Example Response	Response Count
Accurate	"How well it predicts what is trying to be predicted"	9
Consistent	"Will the model change over time"	5
Accurate: Identifies Appropriate patients	"That it never identifies patients who are not appropriate for our intervention. Once it does that, then users will stop finding it useful"	3
Accurate: Across subpopulations	"Consistent outputs across time and is accurate across different subpopulations"	2

TABLE 11 Survey responses to "what does it mean to you for a model to be fair?".

Theme	Example Response	Response Count
Similar Model Performance across demographics	"It doesn't over or under flag patients based on race, ethnicity, age or sex"	6
Similar Model Performance across demographics: Race/Ethnicity	"The model would treat all people the same, regardless of sex or race"	4
Similar Model Performance across demographics: Sex	"Performance is not preferentially high or low based on race, sex, etc."	4
Depends on How Model is Used	"I'm not sure a model is inherently fair or not fair, it seems to me that the way the model is used could be fair or unfair. In one context, being more sensitive for patients of a certain group could be good (fair) for those patients, in another context it could be bad (unfair)."	2
Similar Model Performance across demographics: Age	"To not over or under flag patients based on race, ethnicity, age or sex"	2
Similar Model Performance across demographics: Intersectional	"Outputs are fair across subpopulations and intersectionality"	1
Representative Patient Data	"Was the patient data representative"	1
Considers Socioeconomic Factors	"Takes into account socioeconomic factors, insurance factors"	1

Audit components affecting decision making

Decision makers felt that every component of the audit would affect their decision to deploy the model, including Summary Statistics, Performance, and Subgroup Performance (10/10); and Calibration and Subgroup Calibration (both 9/10). When asked for any other information they would want included in the audit to support their decision on whether to deploy a model (**Supplementary Table S22**), decision makers most commonly responded with **more reliable race data in EHR** (2/10).

Drivers and barriers for audits and AI model use

Decision makers identified **Findings that AI models are not fair** (10/10), **Findings that AI models are not**

reliable (9/10), and **Academic medicine's push toward racial equity** (9/10) as key drivers to making reliability and fairness audits standard practice (Supplementary Table S23). For key barriers, they tended to identify **Poor demographic data quality** (8/10), **Poor data quality** (6/10), and **Lack of data access** (5/10) (Supplementary Table S24).

Decision makers largely saw **Helps triage patients and identify who would benefit the most** (10/10) and **Shared understanding of patients for our whole care team** (9/10) as key advantages of using AI to support their work (Supplementary Table S25). When asked what cons they see in using an AI model to support their work, decision makers tended to respond with **Lack of transparency of the model** (5/10) and **Takes effort to maintain** (4/10) (Supplementary Table S26).

Time and resources required to perform audit

We documented the main tasks, persons performing each task, and estimated time required to perform each task in Supplementary File S1, summarizing in Table 12. Note: we estimated response time per clinician using the median time per surprise question from our decision maker survey responses: 1 min for Primary Care and for Hospital Medicine, and 2 min for Inpatient Oncology.

Averaged across the three settings, we spent 115 h on the audit. Some of the most time-intensive tasks involved processing and analysis of the data (48 person-hours), soliciting clinician labels (24 person-hours), designing and implementing an incentive program to support gathering

TABLE 12 Time and requirements to generate reliability and fairness audits. For further detail, see Supplementary File S1.

	Average across 3 Settings (estimated person-hours)	Primary Care (estimated person – hours)	Inpatient Oncology (estimated person-hours)	Hospital Medicine (estimated person-hours)	Primary Care (date range)	Inpatient Oncology (date range)	Hospital Medicine (date range)
Sample Size Calculation	**15**	25	10	10	8/12/2021–11/8/2021	8/24/2021–11/8/2021	8/24/2021–11/8/2021
Pull Epic Model Predictions	**1**	1	1	1	11/16/2021	6/14/2021	1/31/2022
IRB for Clinician-Patient Linkage in Primary Care	**9**	9	N/A	N/A	12/7/2021–1/14/2022	N/A	N/A
Clinician Label-Gathering: Solicitation	**24**	25	N/A	22	11/19/2021–2/23/2022	N/A	2/15/2022–3/21/2022
Clinician Label-Gathering: Chocolate Incentive	**24**	26	N/A	22	1/26/2022–2/14/2022	N/A	1/26/2022–2/20/2022
Clinician Label-Gathering: Responses	**7**	6	8	7	2/11/2022–3/7/2022	8/15/2021–3/19/2022	2/21/2022–3/22/2022
Clinician Label-Gathering: Recording Responses		3	2	3	2/11/2022–3/7/2022	8/15/2021–3/19/2022	2/21/2022–3/22/2022
Processing & Analysis	**48**	41	58	44	10/31/2021–4/21/2022	11/22/2021–4/21/2022	3/30/2022–4/21/2022
Presentation	**1**	1	2	1	3/21/2022	3/25/2022, 3/29/2022	3/30/2022
Survey	**8**	8	8	8	3/3/2022–4/23/2022	3/3/2022–4/23/2022	3/3/2022–4/23/2022
TOTAL TIME	**115**	145	88	111	8/12/2021–4/23/2022	6/14/2021–4/23/2022	8/24/2021–4/23/2022
TIME OF ITERATION (Code Iterating for Sample Size Calculation & Reliability/Fairness Audit, and Iterating on Presentation)	**40**	45	45	30			
TOTAL TIME WITHOUT ITERATION	**75**	100	43	81			

clinician labels (24 person-hours), and calculating the required sample size (15 person-hours). Notably, the actual responses by the clinicians and recording of responses by the clinicians required less time (9 person-hours), as did designing and implementing the survey (8 person-hours) and presenting to the decision makers (1 person-hours).

Of the 115 h, we classified 40 (35%) of these hours as iteration time – time that JL spent mainly on iterating on writing code (e.g., for calculating required sample sizes and estimating model performance for each subgroup) or drafting presentation material. If we were to do the same study again at this point, presuming we could bypass the iteration time, the audit could likely be done in 75 h (65% of total hours).

In calendar time, the audits were completed 8–10 months from the start, underscoring the need for balancing competing priorities amongst both study designers and participants, building relationships among team members to enable the project, and waiting for clinicians to respond.

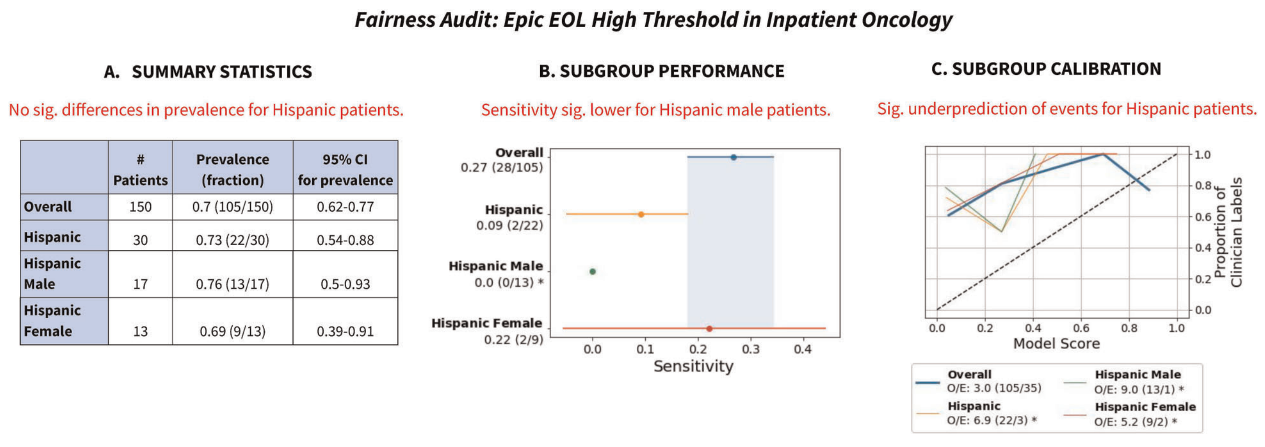
Lastly, we emphasize key requirements in two categories: *stakeholder relationships* and *data access*. On stakeholder relationships, physicians' understanding of the best way to communicate with their colleagues and designing appropriate incentives (e.g., chocolate) were crucial to ensure a high response rate. On data access, there were multiple data sources with different access requirements. Some required healthcare system employees to use their privileged access. For example, KS had to extract Epic model predictions from our EHR for us to perform the audit. Similarly, multiple IT subunits had to coordinate to deliver patient panels for us. Alternatively, other data sources could be accessed using existing data infrastructure. Crucially, our patient demographics and patient visits were already available in a common data format (OMOP-CDM) (44). This allowed iterative querying and refinement to ensure we were pulling the most relevant patients and patient information. Having existing access to a daily hospital census feed and having query access to the hospitalist attending schedules were critical in enabling our hospital medicine clinician labeling workflow (26).

Discussion

We operationalized reliability and fairness audits of predictive models in ACP, with the best attempt to adhere to model reporting guidelines (22). We highlight key insights and themes across audits below and conclude with recommendations for informaticists and decision makers.

Key insights from model fairness audits

We use the Epic EOL High Threshold's performance for Hispanic patients in Inpatient Oncology as an illustrative example (Figure 2) to show the value of reporting summary statistics, subgroup performance and subgroup calibration. (Note: the specific group is Hispanic/Latino patients with Race listed as Other, but we denote them as "Hispanic" patients here for simplicity).



	# Patients	Prevalence (fraction)	95% CI for prevalence
Overall	150	0.7 (105/150)	0.62-0.77
Hispanic	30	0.73 (22/30)	0.54-0.88
Hispanic Male	17	0.76 (13/17)	0.5-0.93
Hispanic Female	13	0.69 (9/13)	0.39-0.91

FIGURE 2
Summary of fairness audit findings, for Hispanic patients for the Epic EOL High Threshold in Inpatient Oncology. (A) Summary statistics revealed no significant difference in prevalence for Hispanic patients. (B) Subgroup performance revealed decreased sensitivity on Hispanic patients and that it was especially low for Hispanic male patients. (C) Subgroup calibration revealed significantly greater miscalibration for Hispanic patients, including for Hispanic female patients and especially for Hispanic male patients. Note: we refer to Hispanic patients with Race listed as Other as Hispanic patients. "Sig." stands for "statistically significant".

First, summary statistics revealed no significant differences in prevalence of the outcome label for Hispanic patients, including after disaggregating by the intersection of race/ethnicity and sex (**Figure 2A**). Assuming no systematic differences in mortality risk or appropriateness of ACP for Hispanic patients vs. Non-Hispanic patients, this reassured us that our surrogate outcome exhibited no obvious signs of bias.

Second, despite insignificant differences in clinician label prevalence, the Epic EOL – High Threshold revealed reduced sensitivity (0.09) for Hispanic patients (**Figure 2B**). The model only flagged 2 of 22 positive patients identified by clinician review. Disaggregation by the intersection of race/ethnicity and sex revealed that the model had significantly reduced sensitivity (0.0) for Hispanic male patients specifically, flagging 0 of 13 positive patients. This demonstrates the value of analyzing model performance for different subgroups (51) and intersectional subgroups (3).

Third, subgroup calibration revealed significant underprediction of events (O/E: 6.9) for Hispanic patients (**Figure 2C**), especially Hispanic male patients (O/E: 9.0). The subgroup calibration shows that the model was systematically giving lower scores to Hispanic patients relative to clinicians, which is potentially linked to the model's lower sensitivity for those groups. Again, this shows how subgroup calibration aids understanding algorithms' impacts on different groups (4).

Differences in the Epic EOL model's sensitivity for Hispanics vs. Non-Hispanics and the model's O/E ratio relative to clinicians for this subgroup also highlights one of the key challenges in using surrogate outcomes (e.g., clinician responses to the surprise question) for reliability and fairness audits. Was the Epic EOL model's sensitivity low for Hispanic Males because it underestimated true risk, or was it that clinicians overestimated risk for those Hispanic Male patients that the model did not flag? Given the consistency of clinician labels across subgroups, we lean toward the former interpretation, but it is impossible to say with certainty in the absence of an objective ground truth label.

Lastly, in all three cases, reporting numerators and denominators put the metrics in context. There were many otherwise seemingly significant results that were marred by low number of patients to calculate the metric for (e.g., for sensitivity, there may be few patients with the positive label). This is especially true for intersectional subgroups that have low representation in the data set (e.g., American Indian or Alaska Native Males).

Consistent themes across audits

Considering the summary statistics of the data sets, there were generally no differences in prevalence of clinician-generated positive labels by sex, race/ethnicity or race/ethnicity and sex. Out of 5 data sets considered, 4 showed either significantly higher prevalence of positive labels for older patients (Age: (70, 80], Age: (80, 90], Age: (90, 100]) or significantly lower prevalence for younger patients (Age: (20, 30], Age: (30, 40]). This is consistent with the older patients having worse prognosis than younger patients and thus was not a cause for concern with respect to label bias. However, it was surprising that for the two Hospital Medicine data sets, there was a higher prevalence of positive labels for non-Hispanic Asian patients (including specifically for those of Male sex) and lower prevalence for Hispanic patients for whom Race was listed as Other (including specifically those of Male sex).

Considering the model performance and calibration, in every setting, all models had high PPV at 0.76 or above; several of our clinicians considered this the most important metric, roughly corresponding to "would a clinician agree if the model flagged a patient?". In Hospital Medicine and Inpatient Oncology, the Epic EOL model at High Threshold tended to flag fewer patients (11%, 21% respectively) than the Stanford HM ACP model (38%, 75%). Meanwhile, the Stanford HM ACP model had higher sensitivity (0.69, 0.89 vs. 0.20, 0.27), and better calibration (O/E 1.5, 1.7) than the Epic EOL model (O/E 2.5, 3.0).

Beyond that, the models often had low sensitivities or PPVs or high rate of underprediction (O/E) for several patient subgroups that had less than 10 patients to compute the metric for in the data set. We emphasize that there is a need to increase representation for these groups so that accurate values can be obtained. Such subgroups include Native Hawaiian or Other Pacific Islander patients, American Indian or Alaska Native patients, Hispanic or Latino patients with race "White" or "Other", and Black or African American patients, among others.

Decision makers overall felt every component of the audit would affect their decision to turn on the model. They most often responded with themes of **Accurate** and **Consistent** for "What does it mean to you for a model to be reliable?". They most often responded with **Similar Model Performance across demographics**, especially for **Race/Ethnicity** and **Sex** for "What does it mean to you for a model to be fair?". The most commonly identified key barriers for making reliability and fairness audits standard practice were **Poor demographic data quality, Poor data quality,** and **Lack of data access.**

Recommendations for informaticists

Invest in checking and improving data validity

Our audit was influenced by multiple unreliable data cascades (52) that hindered our ability to draw decisive conclusions regarding model fairness and reliability. Firstly, it is likely that the race/ethnicity variables were inaccurate, given widespread low concordance with patients' self-identified race/ethnicity found in one of our family medicine clinics (33) and other data sets (34). Thus, a prerequisite for reporting summary statistics and model subgroup performance, as recommended by many model reporting guidelines (9, 11–13, 15, 17, 18, 20, 53), would be better collection of race/ethnicity data. We also again emphasize that race/ethnicity is more a social construct than fixed biological category (32) and the goal of the fairness audit is to understand the demographics of who is represented in data sets and how models impact them. Another data cascade we experienced was large loss of clinician labels after linking these to model predictions and patient demographics (25%–27% for the Epic EOL and 44%–45% for the Stanford HM ACP, in Inpatient Oncology and Hospital Medicine).

Lastly, it is important to verify the validity of source data in detail i.e., *via* manual inspection of the raw data, summary statistics, and metadata for all variables used in the audit. For example, the Sex variable we used from the patient demographic table came from a column called "gender_source_value"; OMOP-CDM documentation (45) clarified "*The Gender domain captures all concepts about the sex of a person, denoting the biological and physiological characteristics. In fact, the Domain (and field in the PERSON table) should probably should be called 'sex' rather than 'gender', as gender refers to behaviors, roles, expectations, and activities in society.*" Relatedly, we found hundreds of visits on a single day for two of the Primary Care providers in the visits table. Our frontline clinicians advised this was likely an artifact given the unrealistic number (AS, WT), so we filtered those two days out.

Perform intersectional analyses

Intersectional analyses proved crucial as they often lended greater clarity to specific subgroups that were being impacted. For example, in Inpatient Oncology, the Epic EOL-High Threshold had low sensitivity (2/22) for Hispanic patients and when disaggregated, specifically had a sensitivity of 0% (0/13) for Hispanic male patients. This would not have been recognized if only looking at sex or race/ethnicity individually. This phenomenon has been discussed in Kimberlé Crenshaw's pioneering intersectionality research to specifically address discrimination against Black women, who often face distinct barriers and challenges relative to White women or Black men (15, 54).

Intersectional subgroup analyses are not difficult to perform, as generating intersectional demographics from one-hot encoded columns only requires performing a logical intersection operation between demographic one-hot encoded columns. However, care must be taken in interpretation of these subgroup analyses as many intersectional subgroups will have poor representation even in large overall sample sizes. Below, we discuss strategies to aid in interpreting results from less frequently represented subgroups.

Contextualize small sample sizes by calculating confidence intervals and reporting metrics as fractions

Small sample sizes of certain subgroups should not be a reason to not consider the subgroups. Proper interpretation of subgroup audit results can be supported by (1) using confidence intervals (e.g., *via* the bootstrap or exact analytical approaches) to appropriately capture sampling variation and (2) reporting metrics with the involved whole numbers (e.g., numerator and denominator, or number of patients) so that if values are extreme, they can be considered in context. For example, several of our bootstrap confidence intervals did not have any width due to there only being one data point from which to resample. [In future work, we would use analytical methods to calculate exact confidence intervals for small sample sizes, such as the Clopper-Pearson interval (47)].

It is especially important to not ignore small sample sizes as doing so can contribute to understudying patient subgroups, especially those that are underrepresented in healthcare data sets due to societal inequities and structural racism. For example, Indigenous peoples have regularly been excluded from COVID-19 data (55) and American Indian and Alaska Native Peoples have often been ignored in data sets due to aggregate analyses (56). Devising sampling strategies in advance to account for known underrepresented populations can help mitigate these issues (e.g., by oversampling underrepresented minorities or increasing sample sizes so that tests for model performance discrepancies between subgroups are adequately powered).

Provider-Patient linkages are necessary data to perform audits using expert-generated labels

Before performing the audit, we did not realize how important it was to be able to generate a list of relevant patients for whom the clinicians would feel comfortable answering the surprise question. Concretely, our clinician annotators felt most comfortable providing labels (the "surprise question") for patients that they had cared for recently. For Primary Care, this required finding recent visits (available in our OMOP-CDM infrastructure) and linking that with patient panels (which we retrieved from business analysts). For Hospital Medicine, this required linking a daily hospital census feed that had assigned treatment teams, with attending- treatment teams. Informatics teams should view clinician-patient linkage as necessary to perform audits in cases where clinician-generated labels are required.

Recommendations for decision makers

Acknowledge limits on data quality for evaluation

Decision makers should recognize the limitations of data quality when performing audits. Race/ethnicity data is likely inaccurate unless proven otherwise given the widespread low concordance with patients' self-identification, as found in our and other data sets (33, 34). Surrogate clinician-generated outcomes used may also be imperfect: our clinician surprise question (a surrogate outcome for appropriateness of an ACP consultation) did not include blinding to the Stanford HM ACP model because it was actively in use as an Epic column as part of the Hospital Medicine SICP implementation. Moreover, while our clinician surprise question generally did not exhibit any obvious differences across ethnicity/race, other studies have found that using surrogate outcomes (e.g., health spending as a proxy for health risk) can exacerbate existing disparities in health (e.g., by estimating that Black patients are at lower health risk because health spending for Black patients has historically been lower than for White patients) (4). Lastly, there were many dropped patients due to lack of an associated model prediction which, if not missing at random, could affect the reliability of our audit.

Require reliability and fairness audits of models before deployment

Our work demonstrates that it is feasible to do thorough reliability and fairness audits of models according to model reporting guidelines, despite low adherence to such guidelines for many deployed models (22). In particular, beyond the usual aggregate model performance metrics, it is straightforward to perform pre-study sample size calculations (41), to report confidence intervals on performance metrics (e.g., using bootstrap sampling), to report summary statistics of the evaluation dataset by subgroup, to share calibration plots and calibration measures, and to do subgroup and intersectional subgroup analyses (3, 15). 90% of our decision makers felt that summary statistics, model performance, model calibration, model subgroup performance and model subgroup calibration affected their decision on whether to turn on the model.

Such audits can be performed by internal organizational teams responsible for deploying predictive models in healthcare (23, 57), with the caveat that internal audits may have limited independence and objectivity (23). Alternatively, regulators may conduct such audits, such as the Food and Drug Administration (FDA)'s proposed Digital Health Software Precertification Program which evaluates real world performance of software as a medical device (58). A more likely scenario is the emergence of community standards (59) that provide consensus guidance on responsible use of AI in Healthcare. We propose that the cost of performing such audits be included in the operating cost of running a care program in a manner similar to how IT costs are currently paid for, with a specific carveout to ensure audits are performed and needed resources are funded.

Enable audits via connecting impacted stakeholders and informaticists

Our decision makers facilitated relationships with their colleagues in Primary Care, Inpatient Oncology and Hospital Medicine that enabled generation of sufficient clinician labels for us to perform our external validation with excellent response rates. This shows the value of interdisciplinary teams and how important it is to honor the trust that comes with personal connections (27, 60, 61). Without this strong relationship, we would have been unable to perform our analysis.

Interpret fairness audits in context of the broader sociotechnical system

Fairness is not solely a property of a model but rather encopmpases the broader sociotechnical system in which people are using a model (62). As one of the decision makers noted, "*I'm not sure a model is inherently fair or not fair, … In one context, being more sensitive for patients of a certain group could be good (fair) for those patients, in another context it could be bad (unfair).*" Furthermore, fairness is not just a mathematical property, but it involves process, is contextual, and can be contested (62). Thus, we note that a fairness audit depicting a model in a favorable light does not by itself *prevent* unfair treatment of patients nor guarantee that use of the model will reduce health disparities.

Conclusion

Despite frequent recommendations by model reporting guidelines, reliability and fairness audits are not often performed for AI models used in health care (21, 22). With respect to reliability, there is a gap in reporting external validation with performance metrics, confidence intervals, and calibration plots. With respect to fairness, there is a gap in reporting summary statistics, subgroup performance and subgroup calibration.

In this work, we audited two AI models, the Epic EOL Index and a Stanford HM ACP model, which were considered for use to support ACP in three care settings: *Primary Care, Inpatient Oncology and Hospital Medicine*. We calculated minimum necessary sample sizes, gathered ground truth labels from clinicians, and merged those labels with model predictions and patient demographics to create the audit data set. In terms of reliability, all models exhibited a PPV of 0.76 or above in all settings, which clinicians identified as the most important metric. In Inpatient Oncology and Hospital Medicine, the Stanford HM ACP model had higher sensitivity and calibration. Meanwhile, the Epic EOL model flagged fewer patients than the Stanford HM ACP model. In terms of fairness, the clinician-generated data set exhibited few differences in prevalence by sex or ethnicity/race. In Primary Care, Inpatient Oncology, and Hospital medicine the Epic EOL model tended to have lower sensitivity in Hispanic/Latino Male patients with Race listed as "Other". The Stanford HM ACP model similarly had low sensitivity for this subgroup in Hospital Medicine but not in Inpatient Oncology.

The audit required 115 person-hours, but every component of the audit was valuable, affecting decision makers' consideration on whether to turn on the models. Key requirements for the audit were (1) stakeholder relationships, which enabled gathering ground truth labels and presenting to decision makers, and (2) data access, especially establishing linkages between providers and patients under their care. For future audits, we recommend recognizing data issues upfront (especially race/ethnicity data), handling small sample sizes by showing confidence intervals and reporting metrics as fractions, and performing intersectional subgroup analyses. Above all, we recommend that decision makers require reliability and fairness audits before using AI models to guide care. With established processes, the 8–10 month calendar time can be compressed to a few weeks given that actual person hours were approximately 3 weeks of effort.

Contribution to the field statement

Artificial intelligence (AI) models developed from electronic health record (EHR) data can be biased and unreliable. Despite multiple guidelines to improve reporting of model fairness and reliability, adherence is difficult given the gap between what guidelines seek and operational feasibility of such reporting. We try to bridge this gap by describing a reliability and fairness audit of AI models that were considered for use to support team-based advance care planning (ACP) in three practice settings: Primary Care, Inpatient Oncology, and Hospital Medicine. We lay out the data gathering processes as well as the design of the reliability and fairness audit, and present results of the audit and decision maker survey. We discuss key lessons learned, how long the audit took to perform, requirements regarding stakeholder relationships and data access, and limitations of the data. Our work may support others in implementing routine reliability and fairness audits of models prior to deployment into a practice setting.

Data availability statement

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author/s.

Ethics statement

The studies involving human participants were reviewed and approved by Stanford University Institutional Review Board. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

Author contributions

JL, AS, SW, ARK, BE, RCL, LS, KR, MFG, SL, WT and NHS contributed to study conception and design. MH, BS and WT engaged key clinical stakeholders. JL performed sample size calculation, advised by YX and SP. KS pulled Epic model predictions. NP and PD supported data access. AS, SW and ARK designed clinician labeling processes for Primary Care, Hospital Medicine and Inpatient Oncology, respectively. AS, SW, ARK, RCL and LS engaged clinicians to perform labeling. AS, SW, ARK, RCL, LS, KR, and others performed labeling. RF implemented the incentive process for clinician labeling, advised by AS, BS and WT. RF and JL solicited and recorded labels for Primary Care and Hospital Medicine, respectively. JL and SL performed data set linking. JL performed quantitative data analysis, advised by AC, SF, BE, SP, YX and NHS. JL designed survey, advised by SW, AC, MS, AG and WT. JL presented results and surveyed AS, SW, ARK, RCL, LS, KR, MFG, SC and WT. JL performed qualitative data analysis, advised by MS and WT. JL wrote first draft of manuscript. JL, AC, SF, and BE wrote sections of manuscript. All authors contributed to the article and approved the submitted version.

Acknowledgments

We would like to thank the primary care faculty at the Stanford Division of Primary Care and Population Health and hospital medicine attending physicians at the Stanford Department of Medicine for their support, time and expertise in generating the labels. We would like to thank Victor Cheng, Henry Nguyen and Ryan Bencharit for support for delivering the Primary Care patient panel. We would like to thank Julian Genkins, Naveed Rabbani, Richard Yoo, and Lance Downing for advising on this work. We would like to thank Randy Nhan, Samantha Lane and Nicholas Kenji Taylor for their poster presentation with Amelia Sattler, and for advising regarding the validity of race/ethnicity in the EHR. We would like to thank Anand Avati and Sehj Kashyap for designing the initial hospital medicine ACP email system, and Anand specifically for supporting with initial suggestions about the sample size analysis.

Conflict of interest

SP is currently employed by Google, with contributions to this work made while at Stanford. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdgth.2022.943768/full#supplementary-material.

References

1. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Intern Med*. (2021) 18:1065–70. doi: 10.1001/jamainternmed.2021.2626

2. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. *J Am Med Inform Assoc*. (2017) 24:1052–61. doi: 10.1093/jamia/ocx030

3. Buolamwini J, Gebru T. *Gender shades: intersectional accuracy disparities in commercial gender classification*. In: SA Friedler, C Wilson, editors. *Proceedings of the 1st conference on fairness, accountability and transparency*. New York, NY, USA: PMLR (2018). p. 77–91. Available at: http://proceedings.mlr.press/v81/buolamwini18a.html

4. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. (2019) 366:447–53. doi: 10.1126/science.aax2342

5. Khetpal V, Shah N. *How a largely untested AI algorithm crept into hundreds of hospitals*. New York, NY, USA: Fast Company (28 May 2021). Available at: https://www.fastcompany.com/90641343/epic-deterioration-index-algorithm-pandemic-concerns (cited 25 Jun 2021).

6. Moons KGM, Kengne AP, Woodward M, Royston P, Vergouwe Y, Altman DG, et al. Risk prediction models: i. Development, internal validation, and assessing the incremental value of a new (bio)marker. *Heart*. (2012) 98:683–90. doi: 10.1136/heartjnl-2011-301246

7. Rivera SC, Liu X, Chan A-W, Denniston AK, Calvert MJ, SPIRIT-AI and CONSORT-AI Working Group. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Br Med J*. (2020) 370:m3210. doi: 10.1136/bmj.m3210

8. Steyerberg EW, Vergouwe Y. Towards better clinical prediction models: seven steps for development and an ABCD for validation. *Eur Heart J*. (2014) 35:1925–31. doi: 10.1093/eurheartj/ehu207

9. Moons KGM, de Groot JAH, Bouwmeester W, Vergouwe Y, Mallett S, Altman DG, et al. Critical appraisal and data extraction for systematic reviews of prediction modelling studies: the CHARMS checklist. *PLoS Med*. (2014) 11:e1001744. doi: 10.1371/journal.pmed.1001744

10. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Br J Surg*. (2015) 102:148–58. doi: 10.1002/bjs.9736

11. Cohen JF, Korevaar DA, Altman DG, Bruns DE, Gatsonis CA, Hooft L, et al. STARD 2015 Guidelines for reporting diagnostic accuracy studies: explanation and elaboration. *BMJ Open*. (2016) 6:e012799. doi: 10.1136/bmjopen-2016-012799

12. Luo W, Phung D, Tran T, Gupta S, Rana S, Karmakar C, et al. Guidelines for developing and reporting machine learning predictive models in biomedical research: a multidisciplinary view. *J Med Internet Res*. (2016) 18:e323. doi: 10.2196/jmir.5870

13. Breck E, Cai S, Nielsen E, Salib M, Sculley D. *The ML test score: a rubric for ML production readiness and technical debt reduction. 2017 IEEE international conference on big data (big data)* (2017). p. 1123–32. doi: 10.1109/BigData.2017.8258038

14. Wolff RF, Moons KGM, Riley RD, Whiting PF, Westwood M, Collins GS, et al. PROBAST: a tool to assess the risk of bias and applicability of prediction model studies. *Ann Intern Med*. (2019) 170:51–8. doi: 10.7326/M18-1376

15. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, et al. *Model cards for model reporting*. In: *Proceedings of the conference on fairness, accountability, and transparency*. New York, NY, USA: Association for Computing Machinery (2019). p. 220–9. doi: 10.1145/3287560.3287596

16. Sendak MP, Gao M, Brajer N, Balu S. Presenting machine learning model information to clinical end users with model facts labels. *NPJ Digit Med*. (2020) 3:41. doi: 10.1038/s41746-020-0253-3

17. Hernandez-Boussard T, Bozkurt S, Ioannidis JPA, Shah NH. MINIMAR (MINimum information for medical AI reporting): developing reporting standards for artificial intelligence in health care. *J Am Med Inform Assoc*. (2020) 27:2011–5. doi: 10.1093/jamia/ocaa088

18. Norgeot B, Quer G, Beaulieu-Jones BK, Torkamani A, Dias R, Gianfrancesco M, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat Med*. (2020) 26:1320–4. doi: 10.1038/s41591-020-1041-y

19. Silcox C, Dentzer S, Bates DW. AI-enabled clinical decision support software: a "trust and value checklist" for clinicians. *NEJM Catalyst*. (2020) 1. doi: 10.1056/cat.20.0212

20. Liu X, The SPIRIT-AI and CONSORT-AI Working Group, Rivera SC, Moher D, Calvert MJ, Denniston AK. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med*. (2020) 370:1364–74. doi: 10.1038/s41591-020-1034-x

21. Bozkurt S, Cahan EM, Seneviratne MG, Sun R, Lossio-Ventura JA, Ioannidis JPA, et al. Reporting of demographic data and representativeness in machine learning models using electronic health records. *J Am Med Inform Assoc*. (2020) 27:1878–84. doi: 10.1093/jamia/ocaa164

22. Lu JH, Callahan A, Patel BS, Morse KE, Dash D, Shah NH. Low adherence to existing model reporting guidelines by commonly used clinical prediction models. *bioRxiv. medRxiv*. (2021). doi: 10.1101/2021.07.21.21260282

23. Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, et al. *Closing the AI accountability gap*. In: *Proceedings of the 2020 conference on fairness, accountability, and transparency* (2020). doi: 10.1145/3351095.3372873

24. Raji D. It's time to develop the tools we need to hold algorithms accountable. In: Mozilla Foundation - It's Time to Develop the Tools We Need to Hold Algorithms Accountable. Mozilla Foundation (2022). Available at: https://foundation.mozilla.org/en/blog/its-time-to-develop-the-tools-we-need-to-hold-algorithms-accountable/ (cited 25 Feb 2022).

25. Li RC, Smith M, Lu J, Avati A, Wang S, Teuteberg WG, et al. Using AI to empower collaborative team workflows: two implementations for advance care planning and care escalation. *NEJM Catalyst*. (2022) 3:CAT.21.0457. doi: 10.1056/cat.21.0457

26. Avati A, Li RC, Smith M, Lu J, Ng A, Shah NH. Empowering team-based advance care planning with artificial intelligence. In: Program for AI In Healthcare at Stanford: Empowering Team-Based Advance Care Planning with Artificial Intelligence (2021). (25 Mar 2021). Available at: https://medium.com/@shahlab/empowering-team-based-advance-care-planning-with-artificial-intelligence-a9edd5294bec

27. Li R, Wang S, Margaret Smith MBA, Grace Hong BA, Anand Avati BS, Jonathan Lu BS, et al. *Leveraging artificial intelligence for a team-based approach to advance care planning*. Society of Hospital Medicine (2021). Available at: https://shmabstracts.org/abstract/leveraging-artificial-intelligence-for-a-team-based-approach-to-advance-care-planning

28. Lett E, Asabor E, Beltrán S, Cannon AM, Arah OA. Conceptualizing, contextualizing, and operationalizing race in quantitative health sciences research. *Ann Fam Med*. (2022) 20:157–63. doi: 10.1370/afm.2792

29. Bailey ZD, Krieger N, Agénor M, Graves J, Linos N, Bassett MT. Structural racism and health inequities in the USA: evidence and interventions. *Lancet*. (2017) 389:1453–63. doi: 10.1016/S0140-6736(17)30569-X

30. Boyd RW, Lindo EG, Weeks LD, McLemore MR. On racism: a new standard for publishing on racial health inequities. *Health Affairs Blog*. (2020) 10:1. doi: 10.1377/hblog20200630.939347

31. Braun L, Fausto-Sterling A, Fullwiley D, Hammonds EM, Nelson A, Quivers W, et al. Racial categories in medical practice: how useful are they? *PLoS Med*. (2007) 4:e271. doi: 10.1371/journal.pmed.0040271

32. Coates T-N. What we mean when we say "race is a social construct.". *Atlantic*. (2013) 15.

33. Randy Nhan BS, Lane S, Barragan L, Valencia J, Sattler A, Taylor NK. *Validating self-identified race/ethnicity at an academic family medicine clinic*. In: *Society of teachers of family medicine 2021 conference on practice & quality improvement* (2021 Sep 13). Available at: https://stfm.org/conferences/1024/sessions/6969

34. Polubriaginof FCG, Ryan P, Salmasian H, Shapiro AW, Perotte A, Safford MM, et al. Challenges with quality of race and ethnicity data in observational databases. *J Am Med Inform Assoc*. (2019) 26:730–6. doi: 10.1093/jamia/ocz113

35. Lake Research Partners Coalition for Compassionate Care of California. Californians' attitudes and experiences with death and dying. In: Final *c*hapter: Californians' *a*ttitudes and *e*xperiences with *d*eath and *d*ying. (9 Feb 2012). Available at: https://www.chcf.org/publication/final-chapter-californians-attitudes-and-experiences-with-death-and-dying/#related-links-and-downloads (cited 25 Mar 2021).

36. Labs A. Serious illness conversation guide. In: *Stanford medicine serious illness care program*. (2020). Available at: https://med.stanford.edu/content/dam/sm/advancecareplanning/documents/Serious_Illness_Conversation_Guide.pdf (cited 22 Apr 2022).

37. Bernacki RE, Block SD. American College of physicians high value care task force. Communication about serious illness care goals: a review and synthesis of best practices. *JAMA Intern Med*. (2014) 174:1994–2003. doi: 10.1001/jamainternmed.2014.5271

38. EPIC. Cognitive computing model brief: End of life care index. (2020 Jan). Available at: https://galaxy.epic.com/?#Browse/page=1!68!95!100039705&from=Galaxy-Redirect

39. Duan T, Anand A, Ding DY, Thai KK, Basu S, Ng A, et al. *Ngboost: natural gradient boosting for probabilistic prediction*. In: *International conference on machine learning*. PMLR. (2020). p. 2690–700. Available at: http://proceedings.mlr.press/v119/duan20a.html

40. Jeremy Orloff JB. Reading for 24: Bootstrap confidence intervals. In: *MIT open course ware: Introduction to probability and statistics*. (2014). Available at: https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2014/resources/mit18_05s14_reading24/ (cited 10 Jan 2022).

41. Riley RD, Debray TPA, Collins GS, Archer L, Ensor J, Smeden M, et al. Minimum sample size for external validation of a clinical prediction model with a binary outcome. *Stat Med*. (2021) 40:4230–51. doi: 10.1002/sim.9025

42. Downar J, Goldman R, Pinto R, Englesakis M, Adhikari NKJ. The "surprise question" for predicting death in seriously ill patients: a systematic review and meta-analysis. *CMAJ*. (2017) 189:E484–93. doi: 10.1503/cmaj.160775

43. White N, Kupeli N, Vickerstaff V, Stone P. How accurate is the "surprise question" at identifying patients at the end of life? A systematic review and meta-analysis. *BMC Med*. (2017) 15:1–14. doi: 10.1186/s12916-017-0907-4

44. Datta S, Posada J, Olson G, Li W, O'Reilly C, Balraj D, et al. A new paradigm for accelerating clinical data science at Stanford Medicine. arXiv [cs.CY]. (2020). Available at: http://arxiv.org/abs/2003.10534

45. Gender Domain and Vocabulary. In: *Observational health data sciences and informatics*. Available at: https://www.ohdsi.org/web/wiki/doku.php?id=documentation:vocabulary:gender (cited 12 Mar 2016).

46. National Institutes of Health Office of Research on Women's Health. Office of Management and Budget (OMB) Standards. In: *Office of management and budget (OMB) standards*. Available at: https://orwh.od.nih.gov/toolkit/other-relevant-federal-policies/OMB-standards (cited 11 May 2022).

47. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. (1934) 26:404–13. doi: 10.2307/2331986

48. An algorithm that predicts deadly infections is often flawed. Available at: https://www.msn.com/en-us/news/technology/an-algorithm-that-predicts-deadly-infections-is-often-flawed/ar-AALh50A (cited 28 Jun 2021).

49. Reps JM, Ryan PB, Rijnbeek PR, Schuemie MJ. Design matters in patient-level prediction: evaluation of a cohort vs. Case-control design when developing predictive models in observational healthcare datasets. *J Big Data*. (2021) 8:1–18. doi: 10.1186/s40537-021-00501-2

50. van den Goorbergh R, van Smeden M, Timmerman D, Van Calster B. The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression. *J Am Med Inform Assoc*. (2022) 29:1525–34. doi: 10.1093/jamia/ocac093

51. Park Y, Hu J, Singh M, Sylla I, Dankwa-Mullan I, Koski E, et al. Comparison of methods to reduce bias from clinical prediction models of postpartum depression. *JAMA Netw Open*. (2021) 4:e213909. doi: 10.1001/jamanetworkopen.2021.3909

52. Sambasivan N, Kapania S, Highfill H, Akrong D, Paritosh P, Aroyo LM. *"Everyone wants to do the model work, not the data work": data cascades in high-stakes AI*. In: *Proceedings of the 2021 CHI conference on human factors in computing systems*. New York, NY, USA: Association for Computing Machinery. (2021). p. 1–15. doi: 10.1145/3411764.3445518

53. CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med*. (2019) 25:1467–8. doi: 10.1038/s41591-019-0603-3

54. Crenshaw K. Demarginalizing the intersection of race and sex: a black feminist critique of antidiscrimination doctrine, feminist theory, and antiracist politics [1989]. In: Katharine TB, editor. *Feminist legal theory*. New York, NY, USA: Routledge (2018). p. 57–80. Available at: https://www.taylorfrancis.com/chapters/edit/10.4324/9780429500480-5/demarginalizing-intersection-race-sex-black-feminist-critique-antidiscrimination-doctrine-feminist-theory-antiracist-politics-1989-kimberle-crenshaw

55. Goodluck K. The erasure of Indigenous people in U.S. COVID-19 data. In: *The erasure of indigenous people in U.S. COVID-19 data*. (2020). Available at: https://www.hcn.org/articles/indigenous-affairs-the-erasure-of-indigenous-people-in-us-covid-19-data (cited 3 May 2022).

56. Huyser KR, Locklear S. Reversing statistical erasure of indigenous peoples. In: M Walter, T Kukutai, AA Gonzales, R Henry, editors. *The Oxford handbook of indigenous sociology*. Oxford University Press (2021). doi: 10.1093/oxfordhb/9780197528778.013.34

57. Kashyap S, Morse KE, Patel B, Shah NH. A survey of extant organizational and computational setups for deploying predictive models in health systems. *J Am Med Inform Assoc*. (2021) 28:2445–50. doi: 10.1093/jamia/ocab154

58. Center for Devices, Radiological Health. Digital Health Software Precertification (Pre-Cert) Program. In: U.S. food and drug administration. FDA. Available at: https://www.fda.gov/medical-devices/digital-health-center-excellence/digital-health-software-precertification-pre-cert-program (cited 27 Jun 2022).

59. CHAI. Available at: https://www.coalitionforhealthai.org/ (cited 2 Jul 2022).

60. Sendak M, Elish MC, Gao M, Futoma J, Ratliff W, Nichols M, et al. *"The human body is a black box": supporting clinical decision-making with deep learning*. In: Mireille H, editor. *Proceedings of the 2020 conference on fairness, accountability, and transparency*. New York, NY, USA: Association for Computing Machinery (2020). p. 99–109. doi: 10.1145/3351095.3372827

61. Elish MC, Watkins EA. *Repairing innovation: a study of integrating AI in clinical care*. Data & Society (2020).

62. Selbst AD, Boyd D, Friedler SA, Venkatasubramanian S, Vertesi J. *Fairness and abstraction in sociotechnical systems*. In: *Proceedings of the conference on fairness, accountability, and transparency*. New York, NY, USA: Association for Computing Machinery (2019). p. 59–68. doi: 10.1145/3287560.3287598

frontiers | Frontiers in Digital Health

TYPE Brief Research Report
PUBLISHED 31 October 2022
DOI 10.3389/fdgth.2022.958663

OPEN ACCESS

EDITED BY
Mark Sendak,
Other, United States

REVIEWED BY
Parisis Gallos,
National and Kapodistrian University of Athens, Greece
Hiral Soni,
Doxy.me, LLC, United States

*CORRESPONDENCE
Collin J. Engstrom
collin.engstrom@winona.edu

SPECIALTY SECTION
This article was submitted to Health Informatics, a section of the journal Frontiers in Digital Health

RECEIVED 31 May 2022
ACCEPTED 27 September 2022
PUBLISHED 31 October 2022

CITATION
Engstrom CJ, Adelaine S, Liao F, Jacobsohn GC and Patterson BW (2022) Operationalizing a real-time scoring model to predict fall risk among older adults in the emergency department.
Front. Digit. Health 4:958663.
doi: 10.3389/fdgth.2022.958663

Operationalizing a real-time scoring model to predict fall risk among older adults in the emergency department

Collin J. Engstrom[1,2]*, Sabrina Adelaine[3], Frank Liao[3], Gwen Costa Jacobsohn[1] and Brian W. Patterson[1,4]

[1]Department of Emergency Medicine, UW-Madison, Madison, WI, United States, [2]Department of Computer Science, Winona State University, Rochester, MN, United States, [3]Department of Enterprise Analytics, UW Health, Madison, WI, United States, [4]Department of Biostatistics and Medical Informatics, UW-Madison, Madison, WI, United States

Predictive models are increasingly being developed and implemented to improve patient care across a variety of clinical scenarios. While a body of literature exists on the development of models using existing data, less focus has been placed on practical operationalization of these models for deployment in real-time production environments. This case-study describes challenges and barriers identified and overcome in such an operationalization for a model aimed at predicting risk of outpatient falls after Emergency Department (ED) visits among older adults. Based on our experience, we provide general principles for translating an EHR-based predictive model from research and reporting environments into real-time operation.

KEYWORDS

falls prevention, EHR, risk stratification, machine learning, AI, precision medicine

Introduction

Predictive models have the potential to transform clinical care by providing clinical decision support, but only when implemented correctly. A large body of literature exists on the development of models using existing data (1–6), and an increasing number of studies have additionally focused on the importance of designing appropriate interfaces to present the output of models to clinicians (7–9). Less focus has been placed on the technical and system challenges of operationalizing these models by running them in clinical environments in which they can function in real time. This case-study describes challenges and barriers we overcame in the use of such a model after it had been created and validated in silico. Based on this experience, we provide general principles for translating an EHR-based predictive model from research and reporting environments into real-time operation.

Case: Preventing falls after ED visits

Falls are the leading traumatic cause of both injury and death among older adults (age ≥ 65 years) (10). Over 3 million patients who have fallen and require medical care present to US emergency departments (EDs) every year (11); however, the ED itself has not traditionally played a major role in outpatient fall prevention (12). In our health system, 34% of patients presenting to the ED for a fall have had at least one ED visit in the prior six months (13), demonstrating a missed opportunity to connect patients with existing clinical interventions to reduce future fall risk.

Our research team has developed and validated an innovative automated screening algorithm that uses machine learning coupled with electronic health record (EHR) data to predict fall risk in the 180 days following an ED visit using retrospective data (14). This algorithm had the promise of identifying older adults at high risk of falling in the 6 months following the ED visit. Furthermore, engaging with experts in human factors engineering and clinicians, the study team designed a workflow and alerts designed to create a system in which the algorithm facilitates screening of older adult patients in the ED and facilitating referral for fall prevention services (15). Fulfilling this promise required successful translation of the predictive screening algorithm to hospital IT systems and clinical care.

Our task was to operationalize a functional model derived from a research dataset into production. Real impact depended on the ability to translate the research model into a corresponding operational model with minimal effects on model performance.

Steps to operationalization

To be successful, we needed to overcome the translational barriers involved in implementing a real-time machine learning model for predicting older adult ED patients at highest risk for a fall event during the following 6 months. In early meetings between the operational and research teams, we identified several issues with the research model which necessitated changes before implementation would be possible. Firstly, some features used in the research model, while theoretically referring to events that happened in the ED, would not be accessible for use in real time during ED visits. This empirical issue is sometimes referred to as "data latency" or "time travel", where the retrospective data set does not appropriately reflect the real-time availability of the features (16, 17). In our case, diagnosis codes referring to the ED visit were added not only by clinicians at the time of the visit but by professional coders several days later. Additionally, based on our data infrastructure, there was a computational and maintenance advantage to simplifying our model type and decreasing the number of features. For this reason, the diagnosis codes were left out of the final machine learning model during the operationalization phase.

Our model was implemented in three stages. Time from initial discussion with operational stakeholders to active deployment to clinic front line staff was a total of 15 months. As shown in **Figure 1**, the overall process can be thought of as three stages, ranging from training and testing on a research dataset in *Stage 1* to a production-side validation in *Stage 2* to a live implementation in *Stage 3*.

Stage 1: Research dataset

The research dataset used for training and testing consisted of 9,687 instances from patient visits to the Emergency Department (ED) over a span of three and a half years. Roughly 725 features relating to vital signs, past diagnoses, and demographics were selected for the modeling process. In the end, six models were chosen based on area under the ROC curve (AUROC). The AUROC performance of these models ranged from 0.72 for logistic regression to 0.78 for forest-based prediction algorithms (18). Prior to moving from the research environment, we trained models using fewer features and were able to maintain performance while paring down to 15 features. Features involving historic diagnosis

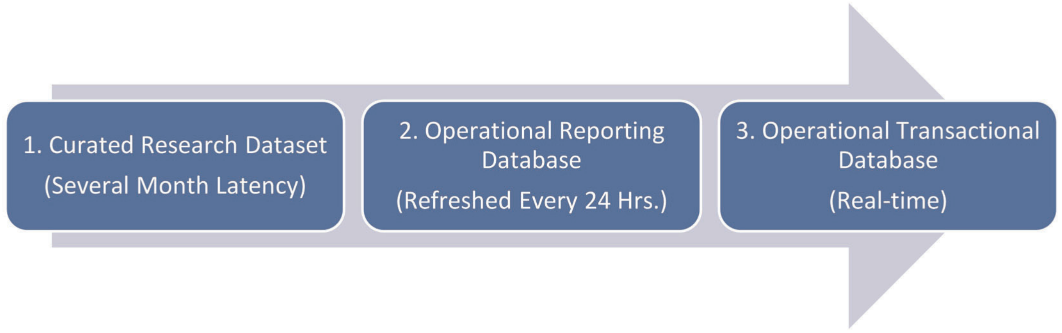

1. Curated Research Dataset
(Several Month Latency)
2. Operational Reporting Database
(Refreshed Every 24 Hrs.)
3. Operational Transactional Database
(Real-time)

FIGURE 1

data, vital signs, and lab values were ultimately left out, as they were not as predictive as initially thought. Height, weight, and age were found to be strong predictors of future falls and were retained, along with those features pertaining to patients' mobility assistance, dementia status, and past occurrences of at-home falls.

Validation in reporting database

Validating our model on the production side during the second stage involved collaboration with the health system's applied data science team. Before moving the model into a real-time scenario, we first validated it using our operational reporting database. This database, while theoretically containing the same information as the research database, required re-querying for features used in our research model to match reporting needs for the production model which would gather data from the electronic health record. This was accomplished by issuing SQL queries to the database one-by-one for features of interest. These features were then fed into the models developed in *Stage 1* and evaluated on the same metrics. This process resulted in an AUROC of 0.69 for a production-ready logistic regression model, a slightly lower, but still acceptable, performance for selecting the most at-risk patients.

Implementation in production transactional database

Planning for final model deployment involved a partnership between physicians, data scientists, computer scientists, health services researchers, and industrial engineers. Ultimately, the features validated in Stage 2 were retrieved from the operational transactional database and forwarded to a model deployed in the cloud, which returns a patient risk score to the EHR. In a separate publication we describe the design of the physician facing interface, an interruptive alert which fires when the returned risk score is above a threshold value. This alert notifies ED clinicians of patients' elevated risk of future fall and facilitates a referral order for our outpatient fall prevention services after an ED visit (15).

Challenges

In moving from *Stage 1* to *Stages 2* and *3*, several unforeseen **feature translation considerations** presented themselves. One of the conveniences of *Stage 1* was the availability of a curated research dataset generated from patient visits. The features from this dataset had been cleaned, however, with some features being removed and new features being added that were derived from others in the dataset. Mapping features to the operational dataset necessitated re-querying features directly from the same data source that would be used in production. Maintenance of the model would require evaluating the consistency of features over time in production data, which would be challenging with so many features. For this reason, the feature set was pared down to include a final production of 15 features. Additionally, ICD data used in the research dataset was not available in real time. For this reason, when moved to the real-time environment we substituted chief complaint data for the ICD data (19).

After these feature-related challenges were overcome, our model was able to compute a risk score for each patient based on the 15 features, all available at time of visit. In our *Stage 1* research, random forest-based models outperformed every other model; however, the difference in practice between these and regression models was minimal. **IT constraints** existed to operationalizing a random forest-based model; as this was the first such model being put into production, there was a strong operational preference for a regression-based model for simplicity of implementation.

From a provider standpoint, this change made sense as well. Providers tend to trust more transparent models that are more explainable (20). Logistic regression is comprised of a linear combination of variables, the importance of which is determined by coefficients that can be interpreted by providers and compared to what they know about falls risk. A desire to ensure we had an **interpretable model** further influenced our choice to pursue regression rather than tree-based models. Also noteworthy was that the physicians interpreting model performance were interested in number needed to treat (NNT) at a given operational threshold (14), a clinical measure that summarizes interventional effectiveness by estimating the number of patients referred to the clinic to prevent a single future fall, rather than AUROC. In summary, our model started as a 700+ feature random forest in the research space but was adapted to a 15-feature regression model for our first operational deployment. This resulted in a small decrease in AUROC and small increase in NNT; however, given the advantages in ease of deployment and maintenance, this was seen as an acceptable tradeoff.

In our research phase, six models (i.e., standard linear and logistic regression, ridge logistic regression, LASSO logistic regression, AdaBoost, and random forests) were tested. For simplicity, the logistic regression model was ultimately chosen as the only one used in production to predict the likelihood of falling six months after leaving the ED. After choosing a model, the **threshold** at which it fired needed to be specified. The clinic to which the intervention referred patients had constraints on the number of patients that they could accommodate each week. This required the model threshold to be adjusted to flag a number of patients commensurate with the operational referral capacity. Our ability to describe

TABLE 1 Guiding considerations: from research to practice.

Consideration for Implementation	Research Design	Operational Adaption
Translational Considerations	725 Features	15 Features
IT Constraints	Tree-Based Models	Logistic Regression
Model Interpretation	Tree-Based Models	Logistic Regression
Communicating model performance and thresholding	Area Under ROC Curve (AUROC) and various NNT thresholds	Adjustable threshold chosen based on NNT and operational capacity
Model Placement in Workflow	Not Considered	Discharge Navigator in Emergency Department

thresholds based on both the number of likely referred patients and the NNT among this group allowed all stakeholders to understand the implications of threshold selection and ongoing adjustment. We have developed a free toolkit which allows calculation of projected NNT at various model thresholds for predictive models, available at www.hipxchange.org/NNT.

Finally, as part of implementing the model in the electronic health record, a point for **model placement in the workflow** had to be chosen so that an interruptive alert would fire, informing the provider to the patient's fall risk. Since all features were available at the time of discharge, this was chosen as the time for the model to run. We describe the design of the alert interface separately (15), but note here that ideal workflow placement of the alert was not achievable, as we were forced to fire the alert at a time when all necessary information to assess patient eligibility was already electronically available in the chart, and further in an area of the chart which was a required portion of the workflow for all discharged patients, to ensure providers would see the alert.

Discussion

Key considerations and questions

While there is increasing recognition that implementation of predictive models requires appropriate validation and governance, the act of moving models from a research platform to operational use presents a unique set of more mundane challenges. In addressing the issues as they arose during the deployment of an EHR-based fall risk prediction model, we identified a series of questions which needed to be addressed. We group these questions below into five domains, summarized in **Table 1** along with examples of our own adaptations in response to these considerations. In future projects, we have found this set of considerations to provide a useful checklist for operationalization.

Feature translation considerations

Do the features that were used to develop the model exist in the context in which the model will be receiving data? Are any features no longer available, or do they change between real-time and retrospective queries? Does the gain in performance from additional features justify the effort needed to create and maintain a more complex model? In our case, review of features from our model revealed that some were not available in real-time. In particular, a diagnosis was thought to be entered only by a physician in the research phase, but most diagnoses are actually entered after a patient visit by professional coders. For this reason, the "diagnosis" feature from the research model was excluded from the production model. Among the features that were available, many did not add enough to our model to justify the additional maintenance and complexity of including them. For example, vital signs and many historical diagnoses were features that were part of the research set but were ultimately left out of the production model for lack of predictive value.

IT constraints

Is it possible for the organization to implement the model? How might model choice be influenced by a healthcare institution's EHR hardware and software? How can models be kept as simple as possible during implementation? In our case there was a preference for a simpler regression-based model for our first attempt at real time prediction to simplify our technical workload, since this time we have iteratively built to more complex model types for other use cases.

Model interpretation

How will model choice impact provider trust? What metrics will such providers use in assessing model viability? In our case, an additional consideration for moving from tree-based to regression models was the ease of communicating model features and operations to our clinical staff.

Communicating model performance and thresholding

What cut-off should be chosen for a model in classifying patients? How should it be chosen based on model performance and clinical scenario? For our clinical scenario, an adjustable threshold based on projected number of patients referred per week by the model, with the resultant performance expressed in NNT, proved a valuable asset in gaining model trust from our referral partners.

Model placement in workflow

When in the user's workflow is sufficient data to run the model entered into the EHR? How quickly will a score need to be calculated in order to be displayed back to the end user in time for action? Is there a distinct electronic trigger that can be used to act on a model score? In our case, responding to these technical considerations significantly impacted our clinical decision support (CDS) design process; in order to collect the required inputs before sending an alert, we were forced to place the tool later in the workflow than our design indicated was optimal.

Conclusion

As machine learning has seen wider uptake in the healthcare setting, there has been an increased need for translating models developed in silico to the bedside. Our team successfully migrated one such model focused on at-home falls risk to our university's emergency department. The process of doing so revealed several challenges which do not fall explicitly within the realm of model development and validation or within the traditional scope of intervention design from a physician workflow perspective.

Ultimately, these challenges were surmountable, but our experience suggests that model operationalization should not be considered a purely technical barrier to implementation but given early consideration when planning an intervention. We hope that the considerations presented here provide guidance for future translation of models into "the wild" and, more generally, bridge the gap that currently exists between research and practice where modeling techniques are concerned.

Data availability statement

The datasets presented in this article are not readily available because **models were created using potentially identifiable patient data used for quality improvement. Due to patient privacy concerns, we are unable to make this data public**. Requests to access the datasets should be directed to **Brian Patterson, bpatter@medicine.wisc.edu**.

Author contributions

CJE and BWP conceived this study. All authors were part of the operational team which performed the technical work reported. CJE and SA drafted the manuscript with significant revisions by BWP and FJL. All authors contributed to the article and approved the submitted version.

Funding

This project was supported by Grant Nos. K08HS024558 and R18HS027735 (Brian W. Patterson) from the Agency for Healthcare Research and Quality (AHRQ). The content is solely the responsibility of the authors and does not necessarily represent the official views of the AHRQ.

Acknowledgments

Authors would like to thank Manish N. Shah, MD, MPH as well as Maureen Smith MD PhD for their advice on development and operationalization of the model.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Dasgupta A, Sun YV, König IR, Bailey-Wilson JE, Malley JD. Brief review of regression-based and machine learning methods in genetic epidemiology: the genetic analysis workshop 17 experience. *Genet Epidemiol.* (2011) 35(S1): S5–S11. doi: 10.1002/gepi.20642

2. Liu C, Che D, Liu X, Song Y. *Applications of machine learning in genomics and systems biology*. Hindawi (2013).

3. Lee EK, Yuan F, Hirsh DA, Mallory MD, Simon HK, editors. *A clinical decision tool for predicting patient care characteristics: patients returning within 72hours in the emergency department. AMIA annual symposium proceedings*; 2012: American Medical Informatics Association.

4. Levin S, Toerper M, Hamrock E, Hinson JS, Barnes S, Gardner H, et al. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. *Ann Emerg Med.* (2018) 71(5):565–74.e2. doi: 10.1016/j.annemergmed.2017.08.005

5. Amalakuhan B, Kiljanek L, Parvathaneni A, Hester M, Cheriyath P, Fischman D. A prediction model for COPD readmissions: catching up, catching our breath, and improving a national problem. *J Community Hosp Intern Med Perspect.* (2012) 2(1):9915. doi: 10.3402/jchimp.v2i1.9915

6. Weiss JC, Natarajan S, Peissig PL, McCarty CA, Page D, editors. *Statistical relational learning to predict primary myocardial infarction from electronic health records. Twenty-Fourth IAAI conference*; 2012.

7. Wiens J, Saria S, Sendak M, Ghassemi M, Liu VX, Doshi-Velez F, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med*. (2019) 25(9):1337–40. doi: 10.1038/s41591-019-0548-6

8. Barda AJ, Horvat CM, Hochheiser H. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC Med Inform Decis Mak*. (2020) 20(1):1–16. doi: 10.1186/s12911-020-01276-x

9. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A, editors. *What clinicians want: contextualizing explainable machine learning for clinical end use. Machine learning for healthcare conference*; 2019: PMLR.

10. Burns E, Kakara R. Deaths from falls among persons aged >/=65 years - United States, 2007–2016. *MMWR Morb Mortal Wkly Rep*. (2018) 67 (18):509–14. doi: 10.15585/mmwr.mm6718a1

11. Haddad YK, Bergen G, Florence C. Estimating the economic burden related to older adult falls by state. *J Public Health Manag Pract*. (2019) 25(2):E17–24. doi: 10.1097/PHH.0000000000000816

12. Carpenter CR, Lo AX. Falling behind? Understanding implementation science in future emergency department management strategies for geriatric fall prevention. *Acad Emerg Med*. (2015) 22(4):478–80. doi: 10.1111/acem.12628

13. Patterson BW, Smith MA, Repplinger MD, Pulia MS, Svenson JE, Kim MK, et al. Using chief complaint in addition to diagnosis codes to identify falls in the emergency department. *J Am Geriatr Soc*. (2017) 65(9):E135–e40. doi: 10.1111/jgs.14982

14. Patterson BW, Engstrom CJ, Sah V, Smith MA, Mendonca EA, Pulia MS, et al. Training and interpreting machine learning algorithms to evaluate fall risk after emergency department visits. *Med Care*. (2019) 57(7):560–6. doi: 10.1097/MLR.0000000000001140

15. Jacobsohn GC, Leaf M, Liao F, Maru AP, Engstrom CJ, Salwei ME, et al. Collaborative design and implementation of a clinical decision support system for automated fall-risk identification and referrals in emergency departments. *Healthcare (Amsterdam, Netherlands)*. (2022) 10(1):100598. doi: 10.1016/j.hjdsi.2021.100598

16. 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(1):144–51. doi: 10.1136/amiajnl-2011-000681

17. Chan KS, Fowles JB, Weiner JP. Review: electronic health records and the reliability and validity of quality measures: a review of the literature. *Med Care Res Rev*. (2010) 67(5):503–27. doi: 10.1177/1077558709359007

18. Patterson BW, Engstrom CJ, Sah V, Smith MA, Mendonça EA, Pulia MS, et al. Training and interpreting machine learning algorithms to eva. *Medical Care*. (2019) 57(7):56–6566. doi: 10.1097/MLR.0000000000001140

19. Patterson BW, Jacobsohn GC, Maru AP, Venkatesh AK, Smith MA, Shah MN, et al. RESEARCHComparing strategies for identifying falls in older adult emergency department visits using EHR data. *J Am Geriatr Soc*. (2020) 68 (12):2965–7. doi: 10.1111/jgs.16831

20. Diprose WK, Buist N, Hua N, Thurier Q, Shand G, Robinson R. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *J Am Med Inform Assoc*. (2020) 27(4):592–600. doi: 10.1093/jamia/ocz229

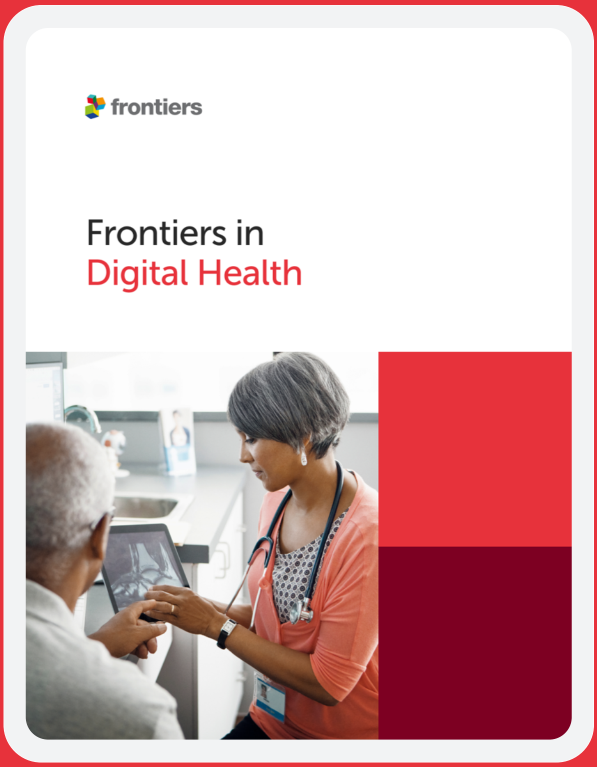
Frontiers in
Digital Health

Explores digital innovation to transform modern healthcare

A multidisciplinary journal that focuses on how we can transform healthcare with innovative digital tools. It provides a forum for an era of health service marked by increased prediction and prevention.

Discover the latest Research Topics

See more →



frontiers

Frontiers in
Digital Health

Frontiers

Avenue du Tribunal-Fédéral 34
1005 Lausanne, Switzerland
frontiersin.org

Contact us

+41 (0)21 510 17 00
frontiersin.org/about/contact