NEJM Catalyst | Innovations in Care Delivery

CASE STUDY

Using AI to Empower Collaborative Team Workflows: Two Implementations for Advance Care Planning and Care Escalation

Ron C. Li, MD, Margaret Smith, MBA, Jonathan Lu, MS, Anand Avati, MS, Samantha Wang, MD, MHS, Winifred G. Teuteberg, MD, Kenny Shum, PhD, Grace Hong, Briththa Seevaratnam, MS, Jerri Westphal, MSN, RN, CNML, Margaret Dougherty, MHA, MS, OTR/L, Pooja Rao, MS, OTR/L, CSRS, Steven Asch, MD, MPH, Steven Lin, MD, Christopher Sharp, MD, Lisa Shieh, MD, PhD, Nigam H. Shah, MBBS, PhD

Vol. 3 No. 4 | April 2022

DOI: 10.1056/CAT.21.0457

To facilitate the development of machine-learning (ML) models in care delivery, which remain poorly understood and executed, Stanford Medicine targeted an effort to address this *implementation gap* at the health system by addressing three key challenges: developing a framework for designing integration of artificial intelligence (AI) into complex health care work systems; identifying and building the teams of people, technologies, and processes to successfully develop and implement AI-enabled systems; and executing in a manner that is sustainable and scalable for the health care enterprise. In this article, the authors describe two pilots of real-world implementations that integrate AI into care delivery: one to improve advance care planning and the other to decrease unplanned escalations of care. While these two implementations used different ML models for different use cases, they shared a set of principles for integrating AI into care delivery. The authors describe how these shared principles were applied to the health system, the barriers and facilitators encountered, and how these experiences guided processes for collaboratively designing and implementing user-centered AI-enabled solutions.

KEY TAKEAWAYS

- » Artificial intelligence (AI) is not the end product, but rather an enabling function in the form of machine-learning (ML)–generated predictions that power a broader set of digital applications, workflows, and human teams (i.e., an *AI-enabled system*).

- » The AI-enabled system must be designed and implemented in a manner that is user centered and driven by pragmatic needs and challenges. The full impact of AI on work systems may emerge via second- or third-order effects that can only be observed once it is implemented in the real-world setting.

- » We observed AI playing an important role in aligning care teams around new collaborative workflows that previously did not exist. The role of AI was not necessarily to provide new information or to replace clinical decision-making, but to function as a *dispassionate mediator* of risk, which mitigated disagreements among team members and empowered nonphysician care team members to drive elements of patient care, such as advance care planning and care escalation due to clinical deterioration.

- » A cross-functional team centered around the development and implementation of the AI system is needed, and it must have expertise not just in ML and data science, but also in clinical informatics, quality improvement, design thinking, enterprise analytics, software and IT applications, and clinical operations.

The Challenge

Despite broad interest in and the promise of artificial intelligence (AI) in health care, there remains a lack of understanding of how AI can meaningfully improve care in complex health care environments. While there has been significant progress in developing machine-learning (ML) models for generating predictions that underlie the *intelligence* that comes with AI, the manner in which this intelligence can be incorporated into health care delivery is still poorly understood and demonstrated.[1,2]

We classify this *implementation gap* at our health system into three categories of challenges: (1) developing a framework for designing integration of AI into complex health care work systems,[3] (2) identifying and building the teams of people, technologies, and processes to successfully develop and implement AI-enabled systems, and (3) executing in a manner that is sustainable and scalable for the health care enterprise.

We describe the application of a shared set of principles for using AI to guide care in two real-world implementations at an academic medical center: one to improve advance care planning (ACP) and the other to decrease unplanned escalations of care for clinically deteriorating patients in the hospital. From these two implementations, we observed an emergent characteristic of how AI was able to mediate improvement, which was to enable new team-based workflows for patient-centered care through empowering nonphysician clinical support services.

The Goal

We sought to demonstrate an approach to using AI in health care that could be operationalized and applied to real-world implementations. A key principle was to view AI not as the solution, but as an enabling function of a broader work system consisting of digital applications, workflows, and human teams.

This approach was applied in two different improvement opportunities at our institution:

1. ACP: conversations that elucidate a patient's values and goals in the course of treating a serious illness are infrequently conducted in the hospital setting. This may lead to care that is not concordant with the patient's goals. The inpatient setting was identified as an opportunity for improving rates of ACP for hospitalized patients.

2. Appropriate care escalation: delayed identification and care of clinically deteriorating hospitalized patients, leading to rapid response teams (RRTs), code events, and unplanned escalations to the ICU that can affect patient morbidity and mortality.

The Execution

Both implementations followed the principle of "designing and building the best possible system for the given improvement opportunity using ML capabilities" rather than "implementing a given ML model." We define a *model* as a function learned from data that map a vector of predictors to a real-valued outcome. *Predictors* are also referred to as *inputs*, *features*, or *variables*; the *outcome* is also referred to as *output*, *label*, or *task*. The following questions guided our execution:

1. What are the improvement goals, metrics, current-state processes, pain points, root causes, and key drivers for improvement?

2. What features of workflows and digital tools would address these key drivers? Which of these can be enabled by AI?

3. What parameters of the ML model (e.g., prediction task, predictive accuracy, and classification threshold) would be required to generate the intelligence that enables these key drivers?

4. How do we select the appropriate ML models that meet these requirements? Do we buy, build, or codevelop? How do we validate and customize the ML models for our improvement needs?

5. How do we design, build, iterate, and implement AI-enabled workflows and applications in a manner that is user centered and problem driven, adaptive to the complexity of the health care environment, and scalable and sustainable for the enterprise?

6. How do we evaluate and scale these implementations?

> *"A key principle was to view AI not as the solution, but as an enabling function of a broader work system consisting of digital applications, workflows, and human teams."*

1. Assessment of the Improvement Opportunity

Both implementations yielded a key insight into how AI can mediate improvement for complex health care settings: by providing an objective benchmark that the entire care team can align around, even if they disagreed with the prediction. This alignment opened an opportunity for physicians and nonphysicians to arrive at a shared mental model of risk that enabled coordination and empowerment of nonphysician team members to take necessary actions.

We arrived at this insight by first trying to understand the problem without any predefined notions of how AI was to be used (or even that AI was needed for the solution). We also used methods from quality improvement to identify concrete improvement goals (increase rates of ACP documentation and decrease rates of unexpected escalations of care from clinical deterioration) and two key drivers that could be enabled by AI.

Key Driver #1: Consistent, Objective Assessment and Communication of Risk

ML models that run continuously and generate risk predictions from patient data in the electronic health record (EHR) can offer an advantage over manual clinician assessments.[4] In both implementations, ML models enabled this key driver by providing consistency and objectivity to the assessment of appropriateness of ACP and need for care escalation.

Key Driver #2: Shared Mental Model of Risk Between Physician and Nonphysician Members of the Care Team

AI can facilitate alignment and coordination by acting as an objective assessor of risk. Patient care in a hospital, while supervised by the attending physician, is highly multidisciplinary, and patients interact with a variety of nonphysician clinical support services, such as nursing, rehabilitation services (physical and occupational therapy), social work, and nutrition. One root cause of process breakdowns for both projects, before implementation, was misalignment of risk perception and lack of coordination between physicians and nonphysician team members in performing needed clinical interventions. We found from stakeholder interviews that in times of disagreement, nonphysician team members frequently did not feel empowered to take action, which may have led to missed opportunities for ACP and early action for clinically deteriorating patients.

2. Conception of the AI-Enabled System

For each implementation of the AI enabled system, we designed a set of digital applications and workflows guided by these two key drivers of achieving consistent, objective assessment of risk and a shared mental model across the care team. The following is a set of common features for both AI-enabled systems:

1. A clinical decision support (CDS) system in the EHR supported by ML model predictions that delivers the same information to both physicians and nonphysician members of the patient care team.

2. A standard, structured workflow that empowers nonphysician care team members to initiate action (within their scope of professional practice) guided by the AI-based CDS system.

3. A shared documentation tool in the EHR linked to the CDS system for each member of the care team to document completion of the workflow (and see each other's documentation).

These features were meant to disrupt the hierarchical, physician-driven workflows that existed for both ACP and care escalation for clinical deterioration and replace them with a more democratized and collaborative system that better leveraged the skills and resources of nonphysician clinical support services (Figure 1).

3. Development and Validation of the ML Models

The above design conceptions guided our selection and refinement of the ML models for each AI-enabled system. Both implementations required our team to think through the following questions:

1. What are the ML model prediction tasks that can enable the previously identified key drivers and system design?

2. What are the runtime requirements of the ML model (e.g., how frequently do predictions need to be generated at deployment)?

3. What is the validation strategy that can best reflect ML model performance for the local implementation setting? How do we select the cohort and outcomes used for the validation?

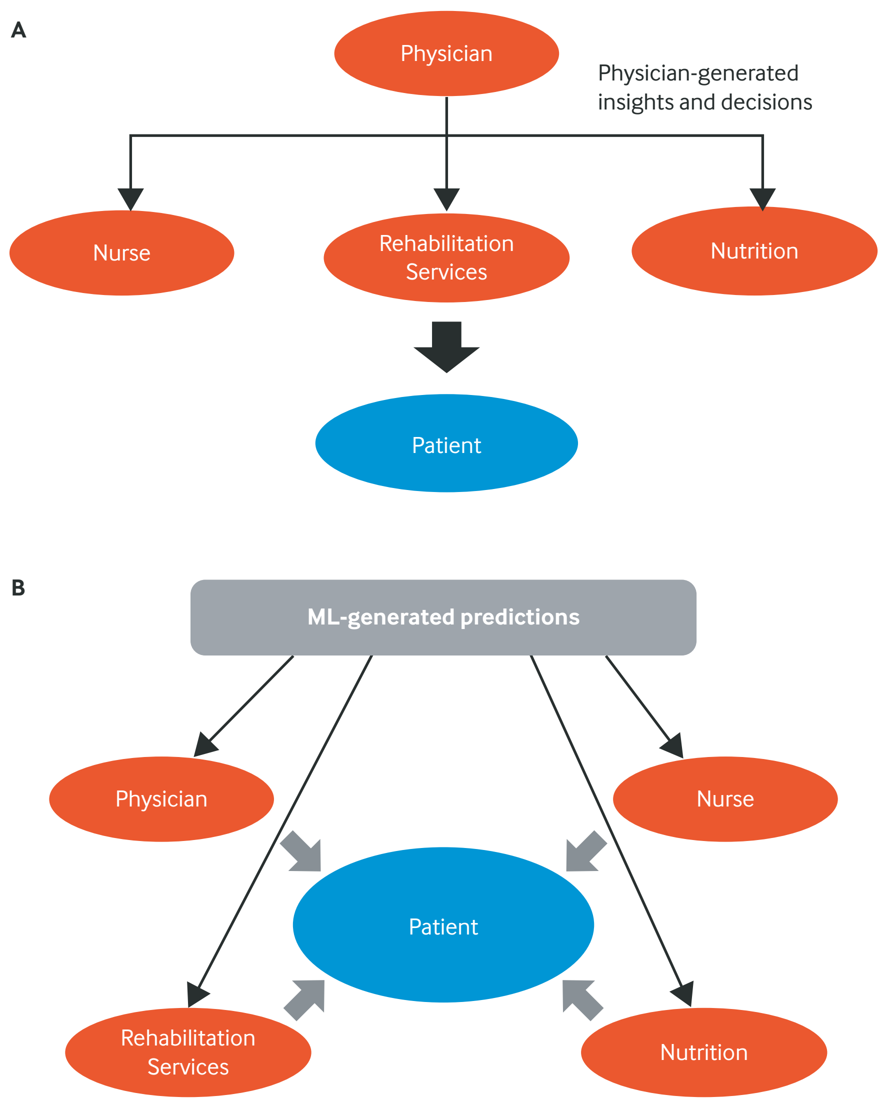
4. How do we select the appropriate classification thresholds for the ML model that can best meet the needs of the system?

> *"We found from stakeholder interviews that in times of disagreement, nonphysician team members frequently did not feel empowered to take action, which may have led to missed opportunities for ACP and early action for clinically deteriorating patients."*

FIGURE 1

Conceptualization of New Collaborative Workflows Enabled by Artificial Intelligence (AI)

For both advance care planning and care escalation, traditional hierarchical workflows (A) involved physicians generating insights and decisions that were then passed down to the rest of the care team and the patient. We envisioned an AI-enabled system (B) in which machine learning (ML)–generated predictions can empower and guide each member of the care team to initiate and carry out decisions in a more democratized and collaborative manner while removing the bottleneck at the level of the physician.



Source: The authors

Implementation #1: Increase Rates of ACP

We aligned on 12-month mortality risk for hospitalized patients as the ML prediction task. Predictions would need to be generated every 24 hours for all admitted patients because the clinical status (and appropriateness of ACP) of hospitalized patients can change over time. We selected a 12-month mortality risk prediction model developed previously by our team that had been validated as an appropriate surrogate for identifying hospitalized patients with serious illness who would benefit from ACP.[5]

The classification threshold was selected so that the model flagged patients in the top 25th percentile of predicted 12-month mortality risk in a cohort of patients discharged from the inpatient general medicine patients at our institution, which reflected the patient population for this implementation. At this threshold, in a larger validation cohort of 5,965 patients who were admitted to our institution, the positive predictive value was 60% (i.e., 60% of patients flagged by the ML model did in fact die within 12 months in the validation cohort). Finally, we estimated the increase in the amount of work in terms of the number of additional patients who would need ACP. In addition, a simulation study was conducted to quantify the achievable net benefit, given that work capacity constraints, as well as patient preferences, often limit follow-up with every flagged case.[6,7]

Implementation #2: Decrease Unplanned Escalation of Care for Clinical Deterioration

To align the care team on the appropriate early interventions, we determined that the ML model needed to identify patients with a high probability of a future clinical deterioration event (e.g., unplanned ICU transfer, RRT, or code), and the predictions would have to be performed early enough to allow for enough time for the care team to respond.[8,9] Predictions would also need to be updated in the EHR to reflect the frequent changes in the patients' clinical status, which enables the first key driver of providing a continuous assessment of risk.

We selected the Deterioration Index (DI), a model available through our EHR vendor, Epic Systems, because of the relative ease of technical integration while meeting most of these requirements. The DI is a logistic regression that is capable of updating predictions on hospitalized patients every 15 minutes using the most recent available clinical data on 31 physiological measures captured in the EHR; the DI tool also shows users the relative contributions of each physiological measure in generating the prediction. This last feature offers the additional benefit of providing a degree of model explainability, which can be useful for helping clinical users align around a shared mental model of risk.[10]

We then performed site-specific validation of the DI on a data set that we derived from a cohort of 6,232 non-ICU patient hospital encounters at our institution using a modified outcome definition that more closely reflected our product requirements: a composite outcome of RRT, code, or ICU transfer within 6 to 18 hours of the prediction.[11] This validation strategy was modified from that of the vendor, which reported model accuracy in predicting the outcomes *without* the 6- to 18-hour time lag; this was thought to not be clinically meaningful because a model predicting an event within 6 hours of the event would not provide sufficient time for a clinical response.

The area under the receiver operating characteristic (AUROC) (which is a performance metric for assessing ML models, in which 0.5 is the worst score and means the model is no better than random chance, and 1.0 is the best) calculated from our validation including these modified definitions was 0.71, which was lower than that reported by the vendor. Given this limited model discrimination and to simplify the model output so that it could be more easily interpreted by the care team, we chose a binary classification threshold (high risk vs. not high risk), which was selected at a cutoff that maximized precision and recall, both of which were 20%. We then validated with a focus group of clinicians that this level of accuracy would indeed be useful (i.e., most agreed they would want to be alerted if their patient had a "1 in 5 chance of experiencing an RRT or ICU transfer within the next 6–18 hours" while acknowledging that "four out of five patients who experience clinical deterioration would *not* be captured by the model"). While the low recall at this threshold (20%) would not make the DI an appropriate comprehensive screening tool for deterioration that would replace existing human-driven screening processes, there was consensus that, at a precision of 20%, it would still be useful to help align mental models and drive the desired physician–nurse team workflows for the patients whom the model *does* flag.

4. Design and Development of AI-Enabled Digital Applications and Workflows

Both implementations included digital applications embedded in the EHR that incorporated ML predictions and enabled shared workflows between physician and nonphysician team members. The EHR applications and workflows were created with two design aims in mind: (1) transparently communicate and align risk across the care team, and (2) promote consistency and collaboration toward patient care.

Communicate and Align Risk Across the Care Team

The following key product features were shared across the two implementations:

- ML predictions had to be translated and displayed into usable information that is simple and avoids confusion that could lead to unintended consequences.
- Information had to be integrated into the clinicians' standard work in the EHR.
- Information had to be displayed transparently to all care team members to facilitate a shared mental model and collaborative work across the care team.

> *"While the low recall at this threshold (20%) would not make the DI an appropriate comprehensive screening tool for deterioration that would replace existing human-driven screening processes, there was consensus that, at a precision of 20%, it would still be useful to help align mental models and drive the desired physician–nurse team workflows for the patients whom the model* does *flag."*

On the basis of these requirements, we developed a shared application design that was used by both implementations: a dedicated column that can be incorporated into EHR patient lists, which are used by both physician and nonphysician care team members as part of standard work. Within the column, patients identified as high risk by the ML are flagged (Figure 2). We decided that it was simpler and more useful for the care team to only see this binary classification result (high risk vs. not high risk) rather than individual numerical model predictions, given that neither model was optimized for calibration and that insight into individual predicted risk values was not necessary for alignment of mental models and workflow.

Because clinical deterioration and care escalation are more acute issues than ACP, we built additional alerting mechanisms in the form of *best practice alerts* in the EHR, as well as interruptive alerts to provider mobile devices for select instances (i.e., those cases in which the system flags the patient as high risk for the first time in the past 24 hours) (Figure 3).

Promote Consistency and Collaboration for Care Delivery

For this second design aim, both implementations shared the following key features:

- Structured workflow shared across the care team for patients flagged by the ML models
- Documentation tools in the EHR that promoted structure, collaboration, and transparency across the care team

We discovered that structure was important for aligning care teams around a collaborative clinical response for flagged patients. A key barrier to the adoption of AI systems in health care that we also observed in our implementation is that clinicians disagree with the ML predictions or believe that the AI system is not telling them anything that they do not already know. In our implementations, the emphasis was less on whether or not the ML predictions were correct; rather, it was that for any given patient flagged by the ML model, physician and nonphysician care team members had to carry out a structured collaborative workflow to build a shared mental model of risk and a collaborative clinical response *regardless of whether there is agreement with the ML prediction.* The role of the AI system was not necessarily to provide new information or to replace clinical decision-making, but to function as a *dispassionate mediator* for facilitating physician and nonphysician collaboration to assess the care plan in light of the new ML-generated information.

To promote consistency in this collaboration, we created the following structured workflows for each implementation.

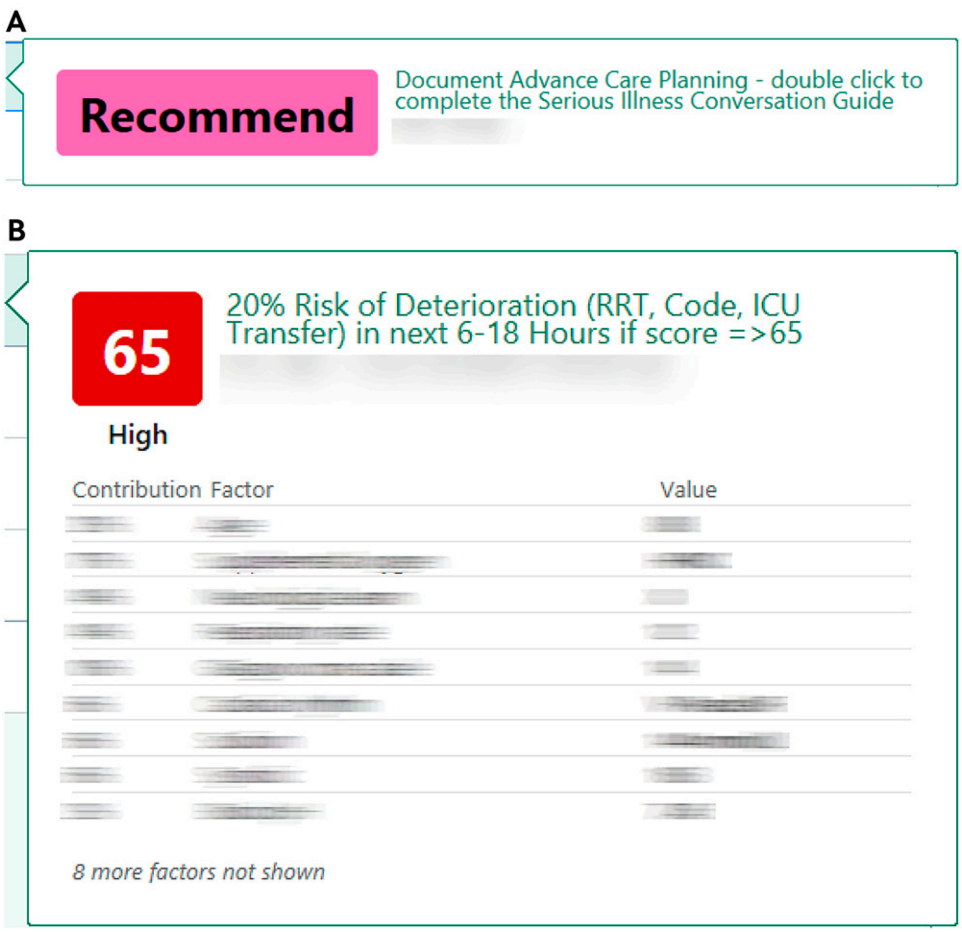
Shared Completion of ACP by Physician and Nonphysician Care Team Members

For patients flagged by the 12-month mortality prediction model, the care team (starting with physicians and occupational therapists and with plans to expand to social worker and clinical nutrition) is asked to conduct ACP using the Serious Illness Conversation Guide (SICG), a

FIGURE 2

Electronic Health Record (EHR) Designs for Communicating Machine-Learning Model Predictions

A patient list column in the EHR was created for each implementation that could be added by both physicians and nonphysician team members to their daily patient lists. Flags were displayed and messages provided when recommended by the machine-learning models. For advance care planning (A), the message was a simple prompt to document advance care planning. For additional evaluation for care escalation (B), the Deterioration Index also included a feature that shows the relative statistical contributions of each variable to the prediction, which provided a particularly helpful context within which clinicians could determine if the model predictions "made sense" (i.e., if the prediction was way off from their clinical judgment, then they could check to see if any of the variables perhaps were derived from either incorrect or outdated data in the EHR). RRT = rapid response team.

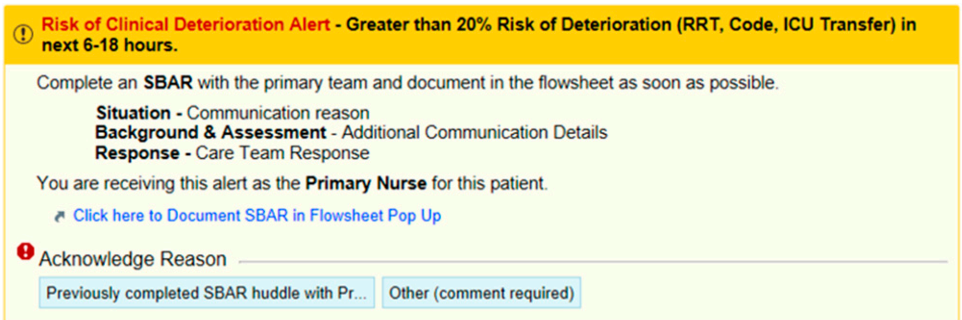


Source: The authors. © 2022 Epic Systems Corporation.

FIGURE 3

Artificial Intelligence–Generated Alerts Sent to Care Teams

A noninterruptive alert is shown on the homepage of the patient's chart if a patient is identified as high risk for clinical deterioration. Additionally, an interruptive alarm is sent to the clinician's phone via the clinical communication mobile application used for patient care (Voalte) for patients who are newly identified as high risk within the previous 24 hours (not shown). RRT = rapid response team, SBAR = Situation, Background, Assessment, Recommendation.



Source: The authors. © 2022 Epic Systems Corporation.

standardized template for ACP using patient-tested language developed by Ariadne Labs (Figure 4).[12,13]

The SICG provides a consistent approach toward ACP that also allows for distributing ownership of different components of the ACP conversation across care team members. In our implementation, the physician is expected to conduct the prognosis component of the conversation, while occupational therapy and nutrition may explore critical abilities that are important near the end of life. Care team members participating in this workflow all underwent training in how to use the SICG.

> *"The role of the AI system was not necessarily to provide new information or to replace clinical decision-making, but to function as a* dispassionate mediator *for facilitating physician and nonphysician collaboration to assess the care plan in light of the new ML-generated information."*

Completion of a Structured Group Huddle for Patients at Risk of Unplanned Care Escalation

The physician and nurse caring for a patient flagged by the ML model were expected to complete a structured huddle — referred to as the *clinical deterioration huddle* — to collaboratively discuss potential reasons for clinical deterioration and next steps (Table 1).

FIGURE 4

The Serious Illness Conversation Guide

The Serious Illness Conversation Guide is a validated template for advance care planning using patient-tested language developed by Ariadne Labs. The Serious Illness Care Program at Stanford Medicine adopted this tool for use by care teams.

Serious Illness Conversation Guide

SETUP

I'd like to talk about what is ahead with your illness and do some thinking in advance about what is important to you so that I can make sure we provide you with the care you want. **Is that okay?**

ASSESS

What is your **understanding** now of where you are with your illness?

How much **information** about what is likely ahead with your illness would you like from me?

SHARE PROGNOSIS

I want to share with you my **understanding** of where things are with your illness.

***Uncertain*:** It can be difficult to predict what will happen with your illness. I **hope** you will continue to live well for a long time, but I'm worried that you could get sick quickly, and I think it is important to prepare for that possibility.

OR

Time: I wish we were not in this situation, but I'm **worried** that time may be as short as ___ (*express as a range, e.g. days to weeks, weeks to months, months to a year*).

OR

Function: I hope that this is not the case, but I'm **worried** that this may be as strong as you will feel, and things are likely to get more difficult.

EXPLORE

What are your most important **goals** if your health situation worsens?

What are your biggest **fears and worries** about the future with your health?

What gives you **strength** as you think about the future with your illness?

What **abilities** are so critical to your life that you can't imagine living without them?

If you become sicker, **how much are you willing to go through** for the possibility of gaining more time?

How much do your **loved ones** know about your priorities and wishes?

CLOSE

I've heard you say ____. Keeping that in mind, and what we know about your illness, I **recommend** that we ____. This will help us make sure that your treatment plans reflect what's important to you.

How does this plan seem to you? **We will do everything we can** to help you through this.

Handoff

To colleague: "I talked with the patient about _____. I learned _____. I think they would benefit from talking with you about _____."

Stanford MEDICINE | Serious Illness Care Program Department of Medicine

http://med.stanford.edu/advancecareplanning

Source: Ariadne Labs, https://www.ariadnelabs.org/serious-illness-care/

NEJM Catalyst (catalyst.nejm.org) © Massachusetts Medical Society

Table 1. SBAR Clinical Deterioration Huddle

Situation	Patient at high risk of clinical deterioration
Background, **A**ssessment	Discuss nursing concerns (primary nurse) and likely reason(s) for clinical deterioration (provider team)
Recommendation	Discuss response to risk of clinical deterioration • Assess aspiration risk • Transfer to high level of care • New orders • Goals-of-care discussion • Family meeting • New consult • ICU provider team consult • Critical care response nurse consult • Other (comment)

Physicians and nurses were expected to complete a structured huddle for flagged patients using the Situation, Background, Assessment, Recommendation (SBAR) format. Source: The authors

The shared checklist format is meant to facilitate consistent incorporation of both physician and nursing perspectives, which we identified as a key driver for improvement.

Integration of AI-Enabled Collaborative Workflows into the EHR

Execution and adherence to these workflows were challenges in both implementations. Clinicians are busy, and their attention is spread out over many complex tasks when caring for patients, so bandwidth for accommodating any new initiative is limited. To address this challenge, we built shared documentation tools in the EHR that incorporated the structure of each AI-enabled workflow and promoted transparency and accountability by each care team member. These tools were all easily accessible by clicking on the patient list flag and alerts generated by the AI system in the EHR.

For the clinical deterioration implementation, tools were incorporated into the EHR to prompt and document the deterioration huddle (Figure 5).

For ACP, the SICG was built into a structured form that allowed care team members to collaboratively conduct ACP and document different sections of the SICG and also for other providers to look back and reference the ACP conversations that have taken place for a patient (Figure 6).

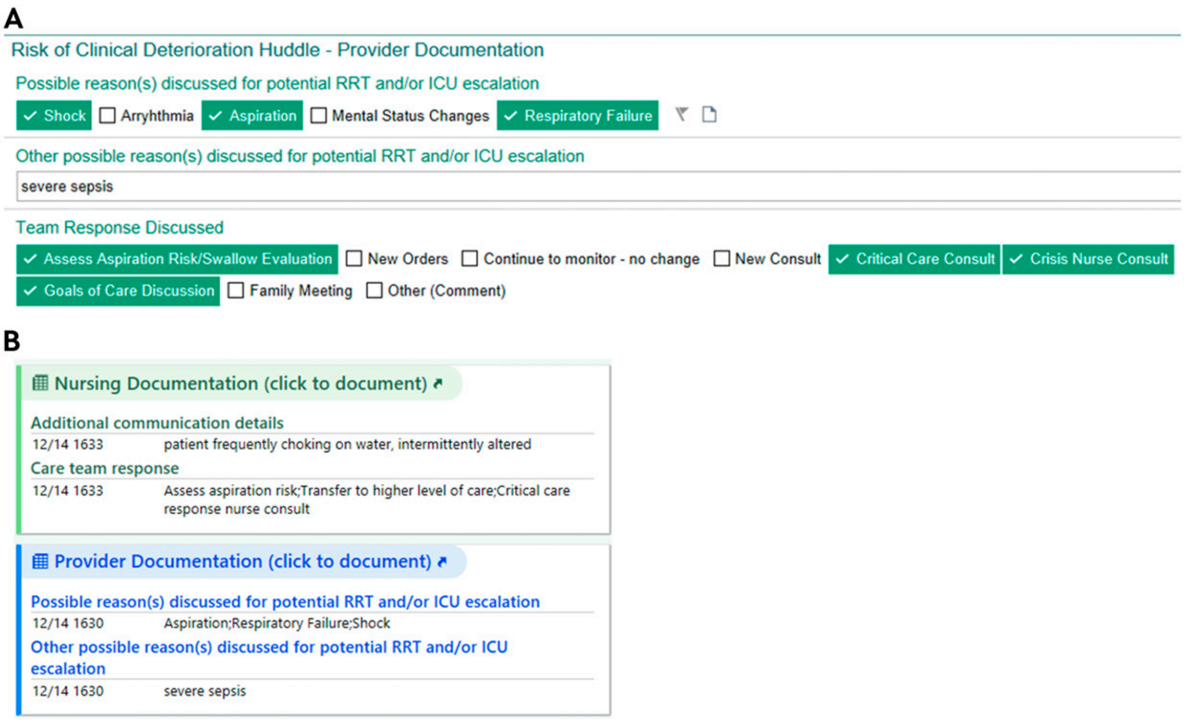
5. Implementation and Testing of Applications and Workflows

Both implementations were initiated and tested on pilot patient care units (the ACP project started in July 2020 and the clinical deterioration project started in January 2021) and followed a *Plan*, *Do*, *Study*, *Act* (PDSA) cycle framework from quality improvement.[14,15] Rapid iteration and testing with deep stakeholder engagement were critical to understanding the barriers and facilitators to implementation. Many design decisions were made after several PDSA cycles that could have surfaced only after real-world experiences and feedback from end users.

FIGURE 5

Collaborative Electronic Health Record (EHR) Documentation Tool for the Clinical Deterioration Huddle

Physicians and nurses complete a checklist for the structured huddle that is embedded into the EHR alert (A). Physician and nursing contributions to the documented huddle are then shown in a report in the patient's chart (B). RRT = rapid response team.



Source: The authors. © 2022 Epic Systems Corporation.

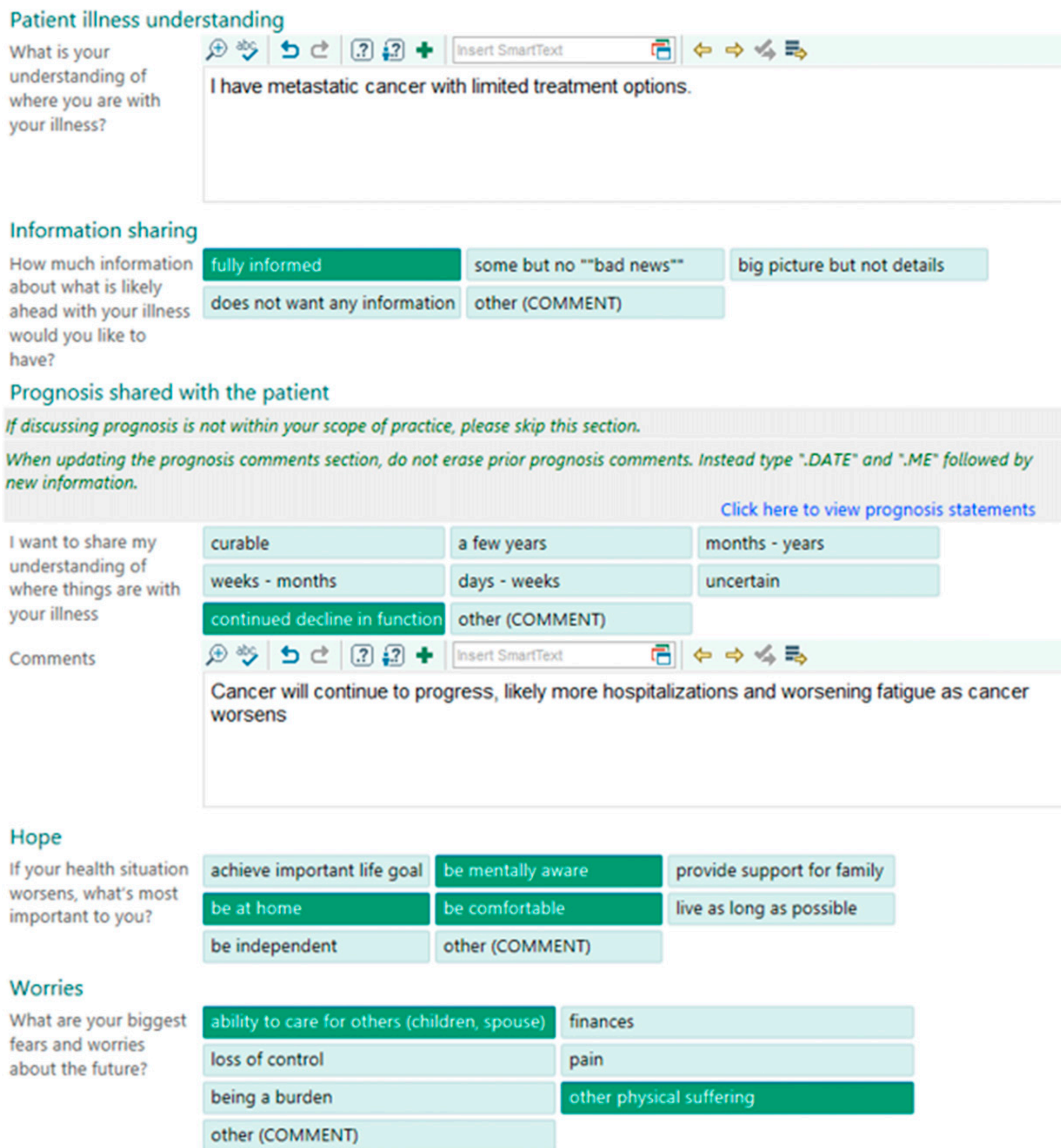
> *"Both implementations have yielded early promising results during the initial pilot phases, as measured by the documented workflow completion rate and interviews with workflow participants."*

For example, for ACP, we initially designed a more coordinated workflow between physicians and nonphysician care team members for flagged patients that included a huddle prior to initiating sections of the SICG. However, on the basis of user feedback regarding bandwidth constraints and the nonacute nature of the ACP relative to other inpatient patient care needs, we instead elected to use a workflow in which any care team member can initiate a section of the SICG for flagged patients (as long as it is within their scope of practice; only physicians were to

FIGURE 6

Collaborative Electronic Health Record (EHR) Documentation Tool for Advance Care Planning

The Serious Illness Conversation Guide was embedded into the EHR as a shared documentation template that can be accessed by double-clicking on the patient list flag. Both physicians and nonphysician care team members can access and edit this documentation template.



Source: The authors. © 2022 Epic Systems Corporation.

NEJM Catalyst (catalyst.nejm.org) © Massachusetts Medical Society discuss the prognosis section); team members would then send a *for-your-information* message to the rest of the team once completed. The AI-generated patient list flag was crucial in facilitating the needed level of alignment for an otherwise decentralized workflow.

6. Integration and Scale into the Health System

As of January 2022, both implementations were nearing the end of their pilot phases, with plans for integration and scale guided by these principles:

Clinical Integration: Ensure that both of the pilot AI-enabled products are sufficiently incorporated into the standard clinical workflows for the care teams to facilitate uptake, beyond just early adopters.

Operational Integration: Connect both implementations to operational units within our institution that are accountable for the metrics and operational goals that these implementations enable.

Technical Integration: Utilize technical infrastructure that can sustainably support the back and front ends of these AI-enabled products at the enterprise level and create a system for monitoring, versioning, and even deimplementing if appropriate.

Metrics

The ACP pilot was implemented for all patients admitted to the general medicine inpatient service, which thus far has included 11,881 total patient hospital encounters since the beginning of the implementation (July 2020) to January 2022 (average of 625 encounters per month), with 2,627 patient encounters flagged by the ML model as candidates for ACP (138 per month; 22% of total encounters).

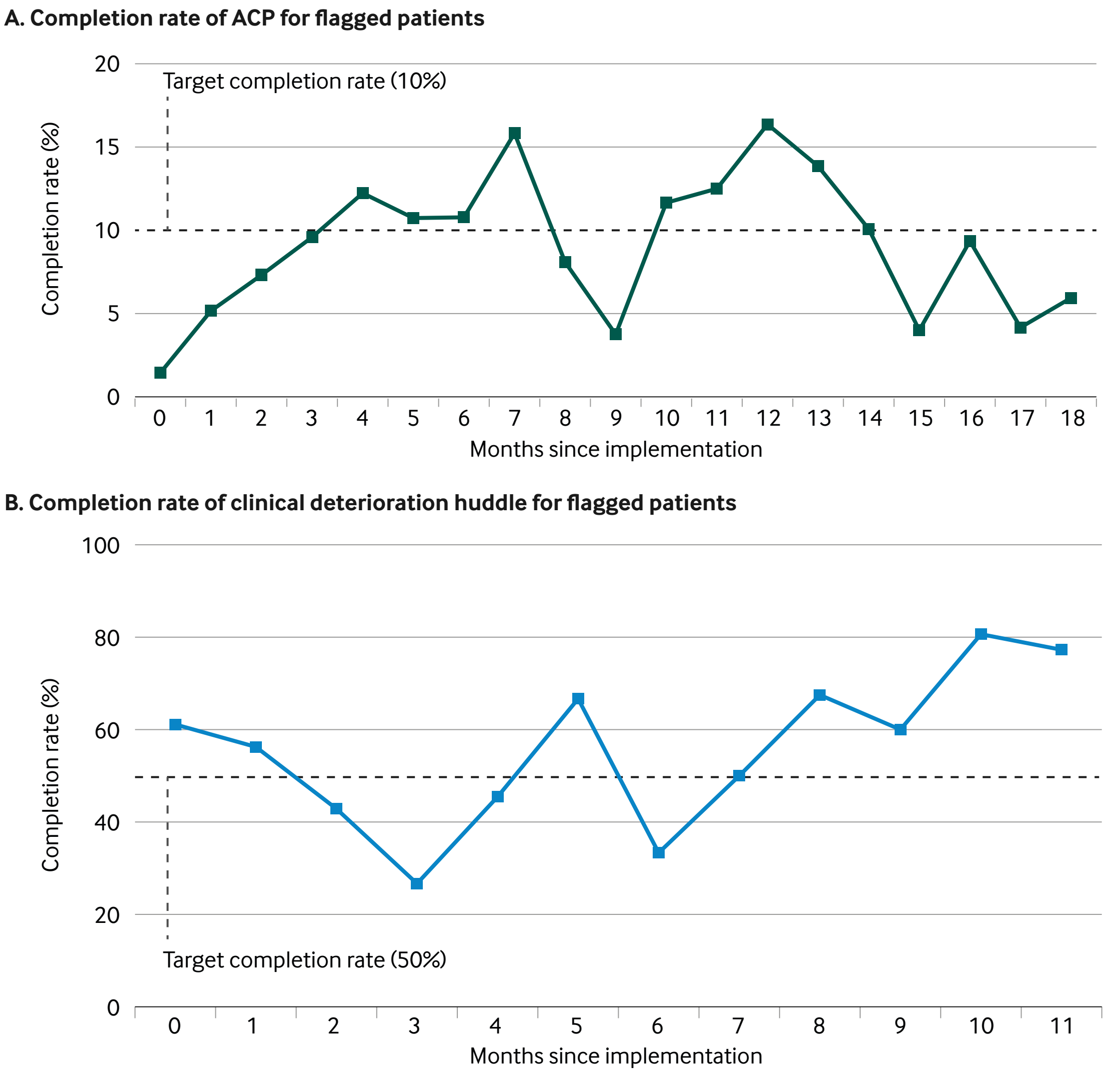
The clinical deterioration pilot was implemented in a stepwise fashion across two different nursing units for general medicine patients, which thus far has included 3,022 total patient encounters since the beginning of the implementation (January 2021) to January 2022 (average of 252 encounters per month), with 313 total patient encounters experiencing at least one flag generated by the DI (average of 21 flags per month; 10.3% of total encounters).

Both implementations have yielded early promising results during the initial pilot phases, as measured by the documented workflow completion rate (Figure 7) and interviews with workflow participants. Target completion rates (number of documented completed workflows/number of flagged encounters) were established for each implementation on the basis of our assessment of clinical appropriateness, estimated number of flagged encounters, time needed to complete each workflow, and capacity of the clinical teams. For ACP, we established a target of 10%, given the higher number of expected flagged encounters, the amount of time needed to complete ACP conversations, and the relatively lower urgency of the intervention for an inpatient encounter. Conversely, we set a higher target (50%) for the clinical deterioration implementation because there are fewer expected flagged encounters, and it was more urgent clinically to complete a

FIGURE 7

Workflow Component Completion Rate

Completion rate of documented advance care planning conversations for patients flagged by the 12-month mortality model is shown in A, and completion rate of documented clinical deterioration huddles is shown in B. Both implementations sustained their target completion rates of 10% and 50%, respectively. There was variability over time due to factors such as varying degrees of work capacity among clinicians and staff and how often the clinical teams deemed the interventions appropriate for the flagged patients. Notably, in the seventh month of the pilot, completion rates for the clinical deterioration huddle increased after an upgrade in the electronic health record documentation tool that improved ease of documentation in July 2021. Note: month 0 for A is July 2020 and for B is January 2021.



Source: The authors

huddle for deteriorating patients; nevertheless, given the precision of 20% (meaning four out of five flagged patients in the validation cohort did not end up needing escalation to the ICU or experienced an RRT or code), we chose not to target 100% workflow completion, because there inevitably will be flagged encounters that clinical teams appropriately decide would not need the full workflow. One limitation to this metric of *documented* workflow completion rate is that it likely underestimates the true rate of workflow completion, because not all completed ACP conversations and clinical deterioration huddles were documented correctly by the clinical teams.

> *“We encountered a number of challenges associated with the implementations, including matters related to time and resources, translating ML predictions into accessible and actionable information, and securing clinician buy-in for the effort. Ultimately, the core of each of these hurdles is rooted in the need to establish trust and confidence in the value of the ML integration.”*

We observed sustained participation from nonphysician care team members for both implementations; of the patient encounters with completed workflows, 100% of completed clinical deterioration huddles included contribution from a nurse, and 42% of completed ACP conversations included contribution from a care team member who was not a physician. In interviews for feedback, many nonphysician members reported that they felt more empowered to leverage their skills to advocate and care for their patients in ways that were not possible before. For example, occupational therapists — who previously were typically not part of ACP conversations (although they had expressed the desire to be so) — reported that they often were able to offer unique perspectives around patient functional goals in the new AI-enabled workflow. One occupational therapist expressed:

“I loved having the [SICG] conversation with a patient today because it really gave me a good understanding of who the patient is as a human being. It was so interesting to see how each person has similar but unique priorities in regard to their medical care and functional goals. The conversation gives us a unique perspective to plan care based on what is important to the patient.”

Nurses have also expressed strong interest in and satisfaction with the AI-enabled clinical deterioration workflow. In a survey of nursing staff from the first of the two pilot nursing units (52 nurses, 30 responded; 57%) on which this implementation was deployed, 96.5% reported that they felt the workflow was adding value to patient care. Additionally, 89.6% indicated that the tool changes the way they care for their patients: charge nurses in the survey reported alternating patient assignments or ratios in anticipation of clinical changes with the flagging patient, and bedside nurses reported they rounded more frequently and/or completed a more in-depth patient assessment on their patients who were flagging.

While nurses have consistently documented completion of the huddles, physician documentation adherence has been minimal. However, survey results shed more light on physician participation and remaining challenges. In a survey among 19 medicine residents participating in the pilot, of whom 17 had at least one patient flagged by the clinical deterioration model, 50% indicated that they take action on the alerts by calling the bedside nurse to huddle, messaging the bedside nurse, or going to the bedside to huddle with the nurse. In addition, 50% indicated that no personal action is taken on the alert; however, 64% said that after receiving an alert, the bedside nurse also reached out to them to discuss the patient's status. When asked about challenges to workflow adherence, 30% of physicians indicated that when they received the alert, they had recently assessed the patient, and, therefore, further action seemed redundant. Providers additionally cited overall workload burden (14%), and disagreement with the model's assessment of risk (14%). Most respondents (68%) reported that they feel either neutral or positive about the overall usefulness of the intervention. These survey results are spurring important conversations and informing key improvements to the overall intervention.

Both implementations will continue to be assessed prospectively with additional quantitative and qualitative outcomes that reflect clinical effectiveness, impact on processes and teams, and success of implementation using implementation science frameworks, such as RE-AIM[16] (reach, effectiveness, adoption, implementation, maintenance) once the pilots have reached a steady state in adoption rate and changes to the workflow (Table 2). Given external factors, such as the Covid-19 pandemic, that have led to multiple unforeseen changes and demands on resources and staffing in the hospital, both of the pilots were extended for an additional 6–8 months to accommodate more PDSA cycles. Examples of additional clinical, process, and implementation metrics for both projects are listed in Table 2.

Hurdles

We encountered a number of challenges associated with the implementations, including matters related to time and resources, translating ML predictions into accessible and actionable information, and securing clinician buy-in for the effort. Ultimately, the core of each of these hurdles is rooted in the need to establish trust and confidence in the value of the ML integration.

Managing Uncertainty Regarding the Value of the ML Component in the Context of Competing Demands for Time and Resources Among Care Teams

Integrating novel workflows into health care is often challenging when there are competing demands for time and resources, especially with the record surges in patient volume our institution has experienced over the course of the implementation period (due to the Covid-19 pandemic and other factors). In particular, workflows involving AI can face a higher barrier to acceptance, because the mechanism triggering the workflow (the ML model) will, by definition, be wrong some percentage of the time (i.e., there is only a certain probability that the patient flagged by the ML model is, indeed, appropriate for the workflow). Additionally, the timing of when these workflows are triggered is critical to adoption and perceived usefulness. This degree of uncertainty can be difficult to understand and accept by clinical teams, especially when other workflows competing for their time and attention are presented with more certainty about expected patient benefit (even if that level of certainty is likely false). For this reason, we designed the EHR

Table 2. Planned Clinical, Process, and Implementation Outcomes for the Two AI-Enabled Systems

	Advance care planning	Clinical deterioration
Clinical effectiveness and process: How did the intervention impact clinical outcomes and processes of care?	• Rates of referral to hospice, palliative care specialists, changes in code status, and hospital readmissions • Frequency and quality of communication and collaboration between physicians and nonphysicians related to patient goals of care	• Rates of overall inpatient mortality, ICU transfers, mortality 24 hours after ICU transfer, conversion of RRTs into codes or ICU transfers • Frequency and quality of communication and collaboration between physicians and nurses related to clinical deterioration
Implementation: How well was the intervention implemented?	• Reach: proportion of flagged patients flagged by the ML model who received ACP • Adoption: proportion of eligible physician and nonphysician providers who participate in the workflow • Implementation fidelity: completion of documented ACP, stratified by provider type and SICG section	• Reach: proportion of patients flagged by the ML model for whom the clinical teams performed a clinical deterioration huddle • Adoption: proportion of eligible physician and nurses who participate in the workflow • Implementation fidelity: completion of documented clinical deterioration huddle, stratified by provider type and components of the huddle

One particular area of focus will be to ascertain the potential impact of the AI system on the level of communication and collaboration between physicians and nonphysician providers related to advance care planning (ACP) and clinical deterioration, which we hypothesize to be a potential downstream effect of the implementations. AI = artificial intelligence, RRT = rapid response team, ML = machine learning, SICG = Serious Illness Conversation Guide. Source: The authors

application builds for both implementations to transparently present the level of statistical uncertainty associated with each ML-generated prediction using user-centered, clinically oriented language so that users can more easily contextualize the relevance of each ML-generated alert.

> “*Implementation efforts will more likely be successful as an improvement opportunity in need of an ML model rather than as an ML model looking for an improvement opportunity.*”

Translating the ML Model Predictions into Interpretable and Actionable Information

Patient care teams need to continuously process large amounts of new information. If that information is ambiguous or not clearly actionable, it is at risk of being misinterpreted, misused, or not used at all. An important lesson we learned is that the ML prediction may not itself be necessarily informative, yet it still plays the important role of aligning clinical teams around a standard set of downstream actions that, on average for flagged patients, may lead to better outcomes. For example, a common piece of feedback we received from clinicians, particularly physicians, for both the ACP and the clinical deterioration implementation was that the model was “not telling [them] anything that [they] don’t already know,” in the sense that they often were already aware that a patient either would benefit from ACP or was at risk of deteriorating. However, despite this prior awareness, physicians often did not actually perform the associated downstream tasks. Therefore, the true value of these AI systems was not necessarily to provide new information, but rather to align the physicians with the rest of the care team around acting on an established workflow.

To incorporate this concept early in each implementation, we pivoted from showing only model predictions to language that specifically outlines the appropriate interpretation and required action. For example, for patients flagged by the DI, nurses (and physicians) received an alert that concretely expressed the nature of the risk and next steps: *"Clinical Deterioration Risk Alert — [insert patient name] is predicted to be at high risk (greater than 20%) of requiring ICU transfer or an RRT in the next 6–18 hours. Connect with the charge nurse and primary team as soon as possible and complete required documentation."*

Building Clinician Trust and Buy-in for the Intervention

The teams employed three strategies to build trust in the models and buy-in for the workflow designed in these implementations. First, site-specific quantitative model validation was conducted for each model, and the results were shared with the clinical stakeholders during the participatory design sessions. Second, clinicians were directly involved in a parallel qualitative model validation process in which they indicated agreement or disagreement with the model predictions. Lastly, the team summarized and shared intervention success stories from early in the pilots to demonstrate patient-level benefit from the intervention. These stories included quotes from staff along with the case details and how the model output informed a different course of action and a favorable outcome.

The Team

In each of the two implementations, a multidisciplinary team consisting of technical, operational, and clinical stakeholders, along with project management and quality improvement support, was convened. More specifically, both project teams included about 15 members: data scientists, clinical informatics, enterprise analytics, nurse managers, frontline nurses, clinical nurse specialists, physicians, project managers, quality improvement experts, and social science researchers. The ACP project additionally included physical therapists, social workers, and dieticians. Engaging all levels of the technical, operational, and clinical stack is a key facilitator of rapid and well-informed decision-making across all phases of the development and implementation of AI-enabled solutions.

Where to Start

The key to starting an implementation project using AI is to pick the right problems to solve that will deliver meaningful improvement for the institution and then build a cross-functional team to develop and integrate the AI-enabled system. This is in contrast to starting with an ML model and trying to figure out how to implement it without a clearly defined problem. Implementation efforts will more likely be successful as an improvement opportunity in need of an ML model rather than as an ML model looking for an improvement opportunity.

Our two implementations went through the following six steps (Table 3) that can be applied to future opportunities.

Table 3. Six Steps for Implementing an AI Workflow Initiative*

Phase	Key components
Assessment of improvement opportunity	• Define the problem statement, improvement targets, and stakeholders • Identify current state gaps and key drivers for improvement that can be enabled by ML
Conceptualization of the AI-enabled system	• Design the components of the newly imagined sociotechnical system enabled by AI that addresses the key drivers
Development and validation of the ML models	• Define appropriate ML prediction tasks • Develop, select, and validate ML models on cohort that reflect the local implementation setting • Determine the appropriate classification thresholds that enable the key drivers and satisfy the work capacity of the team
Design and development of applications and workflows	• Design and build the user-facing digital applications and workflows
Implementation and testing	• Iterate and test the AI-enabled system using the PDSA cycle • Prospectively evaluate pilot implementations
Integration and scale	• Integrate and scale the AI-enabled system into the standard work and processes of the institution

*AI = artificial intelligence, ML = machine learning, PDSA = *Plan, Do, Study, Act.* Source: The authors

Ron C. Li, MD
Clinical Assistant Professor, Stanford University School of Medicine, Stanford, California, USA

Director, Stanford Emerging Applications Lab, Stanford Health Care, Stanford, California, USA

Medical Informatics Director for Digital Health and AI Integration, Stanford Health Care, Stanford, California, USA

Margaret Smith, MBA
Director of Operations, Stanford Healthcare AI Applied Research Team, Stanford University School of Medicine, Stanford, California, USA

Jonathan Lu, MS
Medical Student, Stanford University School of Medicine, Stanford, California, USA

Anand Avati, MS
Graduate Student, Stanford University, Stanford, California, USA

Samantha Wang, MD, MHS
Clinical Assistant Professor, Stanford University School of Medicine, Stanford, California, USA

Assistant Clinical Director, Serious Illness Care Program, Stanford University School of Medicine, Stanford, California, USA

Winifred G. Teuteberg, MD
Clinical Associate Professor, Stanford University School of Medicine, Stanford, California, USA

Clinical Director, Serious Illness Care Program, Stanford University School of Medicine, Stanford, California, USA

Kenny Shum, PhD
Data Scientist, Stanford Health Care, Stanford, California, USA

Grace Hong
Research Assistant, Stanford Healthcare AI Applied Research Team, Stanford University School of Medicine, Stanford, California, USA

Briththa Seevaratnam, MS
Manager, Quality Initiatives and Serious Illness Care Program, Stanford University School of Medicine, Stanford, California, USA

Jerri Westphal, MSN, RN, CNML
Manager, Nurse Informatics, Stanford Health Care, Stanford, California, USA

Former Patient Care Manager, Stanford Health Care, Stanford, California, USA

Margaret Dougherty, MHA, MS, OTR/L
Manager, Rehabilitation Services, Stanford Health Care, Stanford, California, USA

Pooja Rao, MS, OTR/L, CSRS
Occupational Therapist, Stanford Health Care, Stanford, California, USA

Steven Asch, MD, MPH
Professor of Medicine, Stanford University School of Medicine, Stanford, California, USA

Associate Dean for Research, Stanford School of Medicine, Stanford, California, USA

Associate Chief of Staff for Clinical Effectiveness, VA Palo Alto Health Care System, Palo Alto, California, USA

Steven Lin, MD
Clinical Associate Professor, Stanford University School of Medicine, Stanford, California, USA

Founder and Executive Director, Stanford Healthcare AI Applied Research Team, Stanford Medicine, Stanford, California, USA

Vice Chief for Technology Innovation, Division of Primary Care and Population Health, Stanford Medicine, Stanford, California, USA

Christopher Sharp, MD
Clinical Professor of Medicine, Stanford University School of Medicine, Stanford, California, USA

Chief Medical Information Officer, Stanford Health Care, Stanford, California, USA

Lisa Shieh, MD, PhD
Clinical Professor of Medicine, Stanford University School of Medicine, Stanford, California, USA

Associate Chief Quality Officer, Stanford Health Care, Stanford, California, USA

Nigam H. Shah, MBBS, PhD
Professor of Medicine and Biomedical Data Science, Clinical Excellence Research Center; Stanford University School of Medicine, Stanford, California, USA

Chief Data Scientist, Stanford Healthcare, Palo Alto, California, USA

Acknowledgments

We thank the following for their contributions: Melanie Hanson, Former Program Coordinator, Serious Illness Care Program, Stanford University School of Medicine, Stanford, CA; Rebecca Hgah-Guen Fong, Program Coordinator, Quality Initiatives and Serious Illness Care Program, Stanford University School of Medicine, Stanford, CA; Siyun Li, Research Assistant, Stanford University School of Medicine, Stanford, CA; Brian Schenone, RN-BC, CAPM, Lead Informatics Analyst, Stanford Health Care, Stanford, CA; Rob Maclay, Director of Technical Services, Stanford Health Care, Stanford, CA; Rob Gallo, MD, Resident Physician, Stanford Health Care, Stanford, CA; and David Shively, Communications Engineering, Stanford Health Care, Stanford, CA, Thomas Savage, MD, Resident Physician, Stanford Health Care, Stanford, CA, Flora Kechedjian, MSN, RN-BC, Nursing Professional Development Specialist, Stanford Health Care, Stanford, CA, Patrice Duhon, MSN, RN, Director, Epic Optimization and Reporting Patient Care Services, Stanford Health Care, Stanford, CA, Vi Van Nguyen RN, BSN, CMSRN, Epic Optimization Coordinator, Stanford Health Care, Stanford, CA, Gretchen Brown MSN, RN, NEA-BC Associate Chief Nursing Informatics Officer, Stanford Health Care, Stanford, CA, Advait Patil, Student, Stanford University, Stanford, CA.

Disclosures: Ron C. Li, Margaret Smith, Jonathan Lu, Anand Avati, Samantha Wang, Winifred G. Teuteberg, Kenny Shum, Grace Hong, Briththa Seevaratnam, Jerri Westphal, Margaret Dougherty, Pooja Rao, Steven Asch, Steven Lin, Christopher Sharp, Lisa Shieh, and Nigam H. Shah have nothing to disclose.

References

1. Li RC, Asch SM, Shah NH. Developing a delivery science for artificial intelligence in healthcare. NPJ Digit Med 2020;3:107 https://www.nature.com/articles/s41746-020-00318-y https://doi.org/10.1038/s41746-020-00318-y.

2. Smith M, Sattler A, Hong G, Lin S. From code to bedside: implementing artificial intelligence using quality improvement methods. J Gen Intern Med 2021;36:1061-6 https://link.springer.com/article/10.1007%2Fs11606-020-06394-w.

3. Carayon P, Schoofs Hundt A, Karsh BT, et al. Work system design for patient safety: the SEIPS model. Qual Saf Health Care. 2006;15(Suppl 1):i50-8 https://qualitysafety.bmj.com/content/15/suppl_1/i50 https://doi.org/10.1136/qshc.2005.015842.

4. van Smeden M, Reitsma JB, Riley RD, Collins GS, Moons KG. Clinical prediction models: diagnosis versus prognosis. J Clin Epidemiol 2021;132:142-5 https://www.jclinepi.com/article/S0895-4356(21)00013-5/fulltext https://doi.org/10.1016/j.jclinepi.2021.01.009.

5. Avati A, Jung K, Harman S, Downing L, Ng A, Shah NH. Improving palliative care with deep learning. 2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Kansas City, MO, November 13-16, 2017 https://ieeexplore.ieee.org/document/8217669 https://doi.org/10.1109/BIBM.2017.8217669.

6. Jung K, Kashyap S, Avati A, et al. A framework for making predictive models useful in practice. J Am Med Inform Assoc 2021;28:1149-58 https://academic.oup.com/jamia/article/28/6/1149/6045012 https://doi.org/10.1093/jamia/ocaa318.

7. Shah NH, Milstein A, Bagley PhD SC. Making machine learning models clinically useful. JAMA 2019;322:1351-2 https://jamanetwork.com/journals/jama/article-abstract/2748179 https://doi.org/10.1001/jama.2019.10306.

8. Escobar GJ, Turk BJ, Ragins A, et al. Piloting electronic medical record–based early detection of inpatient deterioration in community hospitals. J Hosp Med 2016;11(Suppl 1):S18-24 https://shmpublications.onlinelibrary.wiley.com/journal/15535606 https://doi.org/10.1002/jhm.2652.

9. Kipnis P, Turk BJ, Wulf DA, et al. Development and validation of an electronic medical record-based alert score for detection of inpatient deterioration outside the ICU. J Biomed Inform 2016;64:10-19 https://www.sciencedirect.com/science/article/pii/S1532046416301265?via%3Dihub https://doi.org/10.1016/j.jbi.2016.09.013.

10. McComb S, Simpson V. The concept of shared mental models in healthcare collaboration. J Adv Nurs 2014;70:1479-88 https://onlinelibrary.wiley.com/doi/10.1111/jan.12307 https://doi.org/10.1111/jan.12307.

11. Singh K, Valley TS, Tang S, et al. Evaluating a widely implemented proprietary deterioration index model among hospitalized patients with COVID-19. Ann Am Thorac Soc 2021;18:1129-37 https://www.atsjournals.org/doi/10.1513/AnnalsATS.202006-698OC https://doi.org/10.1513/AnnalsATS.202006-698OC.

12. Ariadne Labs. Serious Illness Care. Accessed November 2, 2021. https://www.ariadnelabs.org/serious-illness-care/.

13. Jacobsen J, Bernacki R, Paladino J. Shifting to serious illness communication. JAMA Netw 2022;327:321-2 https://jamanetwork.com/journals/jama/article-abstract/2788065 https://doi.org/10.1001/jama.2021.23695.

14. Leis JA, Shojania KG. A primer on PDSA: executing plan-do-study-act cycles in practice, not just in name. BMJ Qual Saf 2017;26:572-7 https://qualitysafety.bmj.com/content/26/7/572 https://doi.org/10.1136/bmjqs-2016-006245.

15. Taylor MJ, McNicholas C, Nicolay C, Darzi A, Bell D, Reed JE. Systematic review of the application of the plan-do-study-act method to improve quality in healthcare. BMJ Qual Saf 2014;23:290-8 https://qualitysafety.bmj.com/content/23/4/290 https://doi.org/10.1136/bmjqs-2013-001862.

16. King DK, Glasgow RE, Leeman-Castillo B. Reaiming RE-AIM: using the model to plan, implement, and evaluate the effects of environmental change approaches to enhancing population health. Am J Public Health 2010;100:2076-84 https://ajph.aphapublications.org/doi/10.2105/AJPH.2009.190959 https://doi.org/10.2105/AJPH.2009.190959.