How to validate Machine Learning Models Prior to Deployment: Silent trial protocol for evaluation of real-time models at ICU

Sana Tonekaboni STONEKABONI@CS.TORONTO.EDU
University of Toronto, Vector Institute, The Hospital for Sick Children

Gabriela Morgenshtern MORGENSH@CS.TORONTO.EDU
Azadeh Assadi AZADEH.ASSADI@SICKKIDS.CA
University of Toronto, The Hospital for Sick Children

Aslesha Pokhrel APOKHREL@CS.TORONTO.EDU
University of Toronto, Vector Institute, The Hospital for Sick Children

Xi Huang XI.HUANG1@SICKKIDS.CA
Anand Jayarajan ANANDJ@CS.TORONTO.EDU
University of Toronto, Vector Institute,The Hospital for Sick Children

Robert Greer ROBERT.GREER@SICKKIDS.CA
The Hospital for Sick Children

Gennady Pekhimenko PEKHIMENKO@CS.TORONTO.EDU
University of Toronto, Vector Institute, The Hospital for Sick Children

Melissa McCradden MELISSA.MCCRADDEN@SICKKIDS.CA
The Hospital for Sick Children

Fanny Chevalier FANNY@CS.TORONTO.EDU
University of Toronto, Vector Institute

Mjaye Mazwi MJAYE.MAZWI@SICKKIDS.CA
University of Toronto, The Hospital for Sick Children

Anna Goldenberg ANNA.GOLDENBERG@UTORONTO.CA
University of Toronto, Vector Institute, The Hospital for Sick Children

Abstract

Rigorous evaluation of ML models prior to deployment in hospital settings is critical to ensure utility, performance, and safety. In addition, a guarantee of the usability of such tools requires careful user-centred design and evaluation. Such evaluations can be extra challenging for models that measure unquantified and complex clinical phenomena like the risk of deterioration. This paper introduces a silent trial protocol for evaluating models in real-time in the ICU setting. The trial is designed following principles of formative testing with the goal of evaluating model performance and gathering information that can be used to refine the model to best fit within the intended environment of deployment. We highlight the considerations for a systematic evaluation and explain the design and deployment of the components that enable this trial. We hope that the principles and considerations introduced in this paper can help other researchers validate ML models in their clinical settings.

Data and Code Availability The main contribution of this paper is a novel silent trial protocol; The data and code availability therefore doesn't apply to this work.

1. Introduction

The development of Machine Learning (ML) technologies in healthcare applications are advancing rapidly, but the deployment of such models in real-world clinical settings has not yet become widespread (Sendak et al., 2020a). One of the major reasons for this is the lack of standards to rigorously and comprehensively evaluate such models that would involve the performance accuracy as well as the human usability, safety, and actionability of the tool in practice (Verma et al., 2021; Joyce and Geddes, 2020; McCradden et al., 2022). The difficulty of proper evaluation is exacerbated in the context of ML models that estimate previously unquantified clinical phenomena such as the risk of deterioration. Quantifying such phenomena has significant utility in settings such as the Intensive Care Units (ICUs), where the timely escalation of care for patients experiencing clinical deterioration is challenging and delay in the escalation is often associated with increased risk of in-hospital mortality (Sankey et al., 2016). Due to the lack of well-defined or recorded labels for risk of clinical deterioration, corresponding models are commonly trained and evaluated using proxy labels such as the Modified Early Warning Score (MEWS) for clinical deterioration (Churpek et al., 2016; Kia et al., 2020), or labelled events like septic shock (Henry et al., 2015), cardiopulmonary arrest (CPA) (Tonekaboni et al., 2018; Ong et al., 2012), or unplanned unit transfer (Wellner et al., 2017). However, these proxies fail to capture the heterogeneity of deterioration pathways and so do not accurately estimate the risk of deterioration in real-life settings.

This paper proposes a novel silent trial framework for real-world evaluation and validation of ML models prior to deployment in hospital settings. The trial is designed with the goal of measuring the performance and functionality of complex ML models in clinical settings, assessing the clinical utility of these tools in practice, and gathering information from domain experts for improving the model. The design is based on the principles of formative testing that refers to the exploratory studies that are conducted early in the development cycle of any product and aim to define user profiles, usage modes, functionality specifications, requirements, and workflow (Rubin and Chisnell, 1994). In particular, we present the silent trial with a case study of an ML model that estimates the real-time risk of deterioration in the ICU using streaming physiological signals. Figure 1 shows the road map of this model from design to deployment. After an iterative design cycle, retrospective validation, and model refinement, this model is assessed through the proposed trial. In our silent trial, the model is integrated into the care environment and is tested in real-time. The trial has the objective of both a technical evaluation of the model and gaining a better understanding of how to quantify the currently unquantified "characterization of risk" by domain experts in a complex medical environment. The silent aspect of the trial ensures that, although the predictions are made in real-time, they do not influence care (Wiens et al., 2019). This is to guarantee patient safety and to make sure that the model is being assessed in isolation of its potential impact on the users. But most importantly, this trial is designed to involve clinical experts in the validation process. One reason for this is the complexity of evaluating pathways of clinical deterioration, which are not captured through the Electronic Health Record (EHR) data. The other reason is that we need a clear understanding of users' needs for a successful deployment, and by involving the users in the validation process of the model, we can gather this information and begin to establish trust by engaging users from early stages. Our proposed silent trial enables us to evaluate the model performance, seek to verify assumptions about stakeholders, and also identify high-level function and human factors issues. While these studies are common in medical technology design and evaluation (Rubin and Chisnell, 2008), they are rarely utilized in evaluating clinical ML models (Yang et al., 2019).

Throughout this paper, we highlight the components that make up a rigorous evaluation of a model with respect to its performance, reliability, and usability. We describe the considerations necessary for conducting a validation trial and the technical components required to implement this trial in a hospital setting. Namely, we outline details of a deployment pipeline that integrates the ML model into the care environment and a user interface (UI) that enables the validation of the model. The deployment pipeline for our case study is a streaming data engine designed to integrate the ML model with the data collected at the patient bedside to enable real-time inference. The UI provides an interactive interface that allows evaluators to review the model predictions in real-time and gather expert labels that can be used for model refinement. To the best of our knowledge, our silent trial is the first interactive trial designed to assess a model's utility in a clinical setting using principles of formative testing. We welcome reproductions of this methodology by the community and hope that trials of this sort will help decision-makers better assess the usability of various ML tools in their clinical setting.

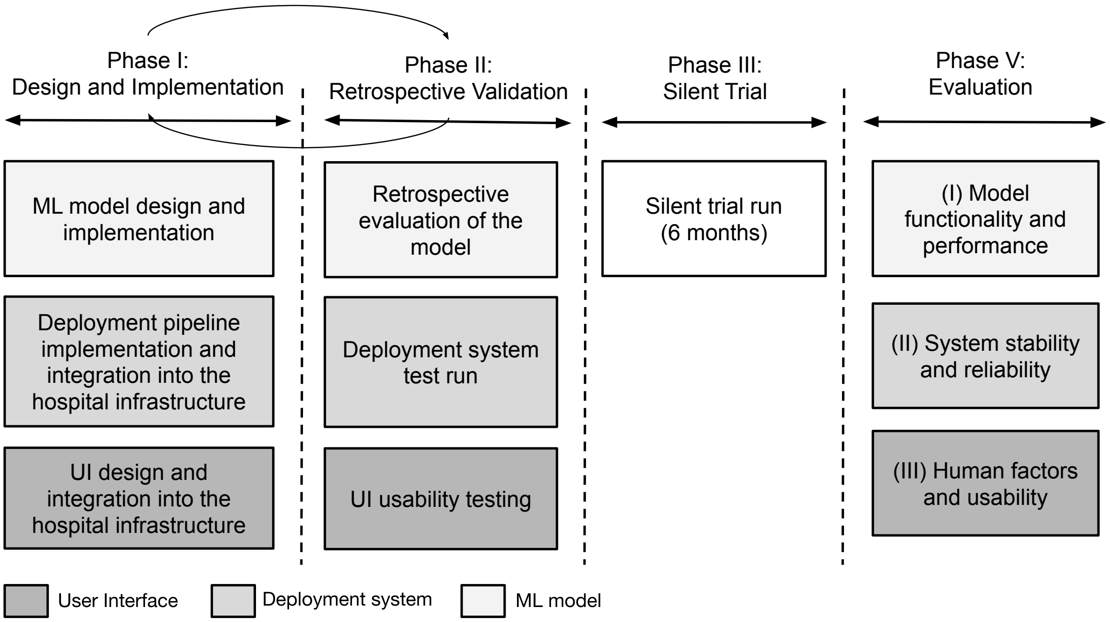

Phase I:
Design and Implementation
Phase II:
Retrospective Validation
Phase III:
Silent Trial
Phase V:
Evaluation
ML model design and implementation
Deployment pipeline implementation and integration into the hospital infrastructure
UI design and integration into the hospital infrastructure
Retrospective evaluation of the model
Deployment system test run
UI usability testing
Silent trial run (6 months)
(I) Model functionality and performance
(II) System stability and reliability
(III) Human factors and usability
User Interface
Deployment system
ML model

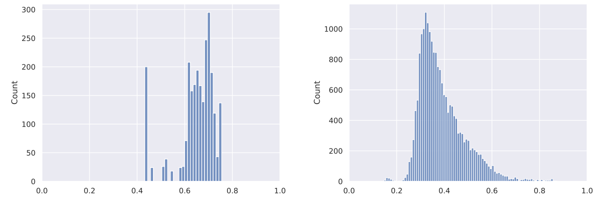
Figure 1: Overview of the path to deployment for the risk predictor model presented in our case study.

2. Background

This section presents the background of our case study for the silent trial. We introduce the ML model of our study, the unique pipeline developed to integrate it in the hospital, and the user interface designed to visualize the predictions in real-time. All these components have gone through cycles of offline evaluations and design prior to the initiation of the silent trial, as described in Figure 1 (Phase I and II).

We validate a deep learning model designed to predict the probability risk of clinical deterioration in ICU patients using high-frequency physiological signals. Since the health condition of patients in the ICU can change rapidly, models that can make real-time estimates of the risk of clinical deterioration have significant utility. Such models can help with the timely escalation of care and ultimately reduce the risk of in-hospital mortality (Sankey et al., 2016). Our model was initially trained using labels for Cardio Pulmonary Arrest (CPA) as high-risk events, serving as proxy labels for deterioration. More details on the model are presented in the Appendix A.1. Patients deteriorate for a variety of reasons, and this heterogeneity in the deterioration pathway (Blackwell et al., 2020; Erez et al., 2021; Rusin et al., 2016) poses a significant challenge in the evaluation of risk models. During the retrospective validation phase of the model using offline data (Figure 1), we measure the predictability of the model in identifying CPA events and look for statistically significant differences in risk estimates for different high-risk and low-risk populations (Figure 2). However, none of these approaches are sufficient for assessing the true clinical accuracy (i.e. ecological validity) of the model in a live environment (McCradden et al., 2020). For instance, some of the false positive predictions of the model might be associated with high-risk events that were detected and prevented by the clinicians, but without any label for such interventions, we would not be able to detect them.

As a result, in order to thoroughly evaluate the usability of our model in a clinical setting and to understand the variety of risk factors that it captures, we need to test the model in the clinical environment and with clinical experts involved. Integrating an ML model into a care environment requires: 1) a deployment system for integrating the ML model with the data collected at the patient bedside, enabling real-time inference; 2) a user interface (UI) to display the real-time risk predictions of the model to clinical users. During multiple iterations of design (Phase I and II in Figure 1) these components are refined to smooth out any friction present in the us-



(*a*) Distribution of risk estimates for individuals prior to ECMO

(*b*) Distribution of risk estimates for all individuals in the ICU

Figure 2: Distribution of the risk estimator model across 2 different patient populations, representing high risk individuals and control. By comparing the two graphs, we can see that the distribution of risk estimates for the two cohorts varies significantly. Note that the estimates are predictions of the model before individuals are put on ECMO.

age of the tool in the clinical environment. Through *usability testing*, the UI is evaluated among the relevant user groups in an offline setting to understand better how the clinical users interact with the tool. More details on the usability testing are provided in Appendix A.2. In sections 4.2.1 and 4.2.2 we provide a more detailed description of each of these components and explain their role in the silent trial.

3. Silent trial protocol

The silent trial protocol is designed to evaluate (in real-time) both the performance of ML models in a clinical setting as well as the utility of these tools in practice. In particular, our trial can be used to validate models that predict clinical events for which there are no well-defined labels with the help of subject matter experts. We describe in detail the aspects of the model that are validated through the silent trial and give specifics on the design and deployment of all the components that enable this validation study (Phase (III) in Figure 1).

3.1. Goals and objectives: What to evaluate during the silent trial?

Below are the three categories of validations we achieve through our silent trial: I) Model functionality and performance, II) System stability and reliability, and III) Human factors concerns.

(I) Model functionality and performance: As a first step, we need to quantify the performance of the ML model in the clinical setting. In new environments, the model must deal with potential data shifts, missing measurements, and corruption or noise on its input. To ensure adequate performance, we need to measure the following:

- Performance metrics: We need to define appropriate metrics (depending on the ML model and the hospital setting) and quantify the performance of the model. These can be either measurable metrics (e.g. False Positive Rate (FPR), precision, accuracy, etc.), or metrics that are defined based on the outcome of using the model (e.g. the number of prevented deaths). In our study, we compare the model's risk prediction with clinician experts' risk assessment of patients.

- Failure modes: ML models can fail under different conditions, related to the quality of the input data or to the complexity of the clinical state. During the trial, we collect annotations from clinical experts through the UI. These annotations capture characteristics of the patient that justify and explain the clinician's assessment of risk. We use these to identify reasons for failure in cases where there is a high discrepancy between the model's prediction and the clinical expert's estimate of risk.

- Bias: Model performance discrepancies across different patient populations can have significant health implications for vulnerable groups, potentially worsening health inequalities (Gianfrancesco et al., 2018; Parikh et al., 2019a; Obermeyer et al., 2019). We should measure the bias in performance across different patient population subgroups in a silent trial. Note that the choice of sub-populations depends on the type of demographic data collected at different institutions. In our case study, we compare the performance of the model for different age groups and patient's first language.

- Data shift: Most ML models suffer from lack of performance robustness, caused by shift in patient population demographics, policy changes across *different institutions* (Subbaswamy and Saria, 2020), or policy changes *over time* (Davis et al., 2017; Nestor et al., 2019). In our case study, we measure the impact of potential data shift over time on the performance of the model.

(II) System stability and reliability: An ML model works hand in hand with the deployment infrastructure, and therefore its performance is also highly dependent on the stability and reliability of the deployment system. During our silent trial, we therefore need to ensure the following:

- Consistency and correctness: Physiological signals collected from the patients are susceptible to delayed, reordered, or missing data points. This unpredictable nature of the raw real-time data streams often causes ML models to make wrong predictions. Ensuring data consistency and correctness is crucial to achieving a robust deployment pipeline. We perform such validations by replaying signal streams from historical datasets, comparing the intermediate results and final output generated by the pipeline to an existing Python-based implementation.

- Availability: Deployed models process the collected physiological signals to continuously monitor patients. It is important to minimize the down-time of a model, to avoid the chance of missing critical clinical events. We keep a log of system failures occurred over the trial period, and try to identify the common causes for failures (e.g. data corruption, load surge, or system malfunction).

- Scalability and efficiency: Deployment pipelines need to process enormous amounts of data under the limited computing infrastructure available within hospitals. Therefore, the pipelines need to be highly scalable, to generate timely predictions, and to efficiently utilize all available hardware resources. To estimate the performance of the pipeline, we measure the data processing throughput and latency of the pipeline, on both single-thread and multi-core setups.

(III) Human factors and usability: Participatory design (PD) plays an important role in the design process of our proposed silent trial, ensuring that end-users are empowered and included in the development of all artifacts involved. An effective implementation plan will empower the right role at the right time of the decision making process (H. Gyldenkaerne et al., 2020). This challenge lends itself nicely to the existing methodology within PD. The goal is not just to understand an activity empirically but to allow stakeholders the opportunity to co-interpret the research (Spinuzzi, 2005). This co-interpretation is a vital component of design work in healthcare and especially vital in deploying ML models to the clinical setting. Without total understanding and participation from the stakeholders at every stage of the project, we cannot have a total understanding of which assumptions are reasonable to make and where innovation should occur to best fulfil stakeholder needs. Thus, to deploy an ML model such as the one proposed here, we must work to understand user needs, and begin to establish trust by involving the users in the validation process of the model. We assess the following through the silent trial:

- Impact of visualization on clinical user: We want to assess whether user's risk assessment would be influenced by being exposed to the model's prediction and if so, we would like to understand the circumstance of this influence. Additionally, we want to confirm that the interface is effective in performing the workflow for which it is designed.

- Stakeholder empowerment: It is important for us to understand how the interface's workflow impacts clinicians' understanding of the model's accuracy within a clinical context. During our usability study, we ask clinicians to assess whether the current design is effective in allowing them to communicate their opinion on the model predictions with sufficient detail and usability.

4. Silent trial case study

4.1. Participants and environment

The particulars of the environment should be taken into account when designing the silent trial protocol. In our environment, we are clinically validating the longitudinal risk estimates of deterioration for all patients in a tertiary pediatric ICU with 42 beds. The intensive care unit is a fast paced, high acuity, and rapidly evolving environment. Patients in the ICU are critically ill, often have marginal hemodynamics, and can quickly deteriorate. Clinical decisions are made rapidly and repeatedly based on evolving patient condition and new clinical findings, by a multitude of different clinicians with a variety of training, expertise, and roles. A wide variety of life sustaining technologies, such as ventilators and extracorporeal membrane oxygenation (ECMO) machines, are also routinely utilized in the ICU, each relying on the expertise of various clinical roles to operate.

To evaluate the model and its user interface, we will be leveraging the expertise of medical trainees (evaluators) who work in our study's ICU environment, but are not directly responsible for patient care while participating in the study. This is to ensure the model's prediction, and the study as a whole, bear no impact on the clinical care of patients during our silent trial. As part of their role, these clinicians are intended to be providing risk surveillance in the unit. They evaluate patients and determine a risk which they use to decide on the next steps to prevent clinical deterioration (Harrigan and Morgenshtern et al., 2021). Therefore, the trial is a way of quantifying and recording the task they already perform to minimize the additional burden on the healthcare provider team. These clinicians are recruited through a town hall, during which they will be introduced to the model, its user interface, and the objectives of the study. All participant questions and feedback regarding the study's methodology and objectives will be addressed before proceeding with the voluntary recruitment of clinicians for the silent trial period. This step aims to ensure stakeholders are acknowledged in the process of validating the model within their work space, and will uncover any last mistaken assumptions about, or areas of mistrust in, the study prior to its execution.

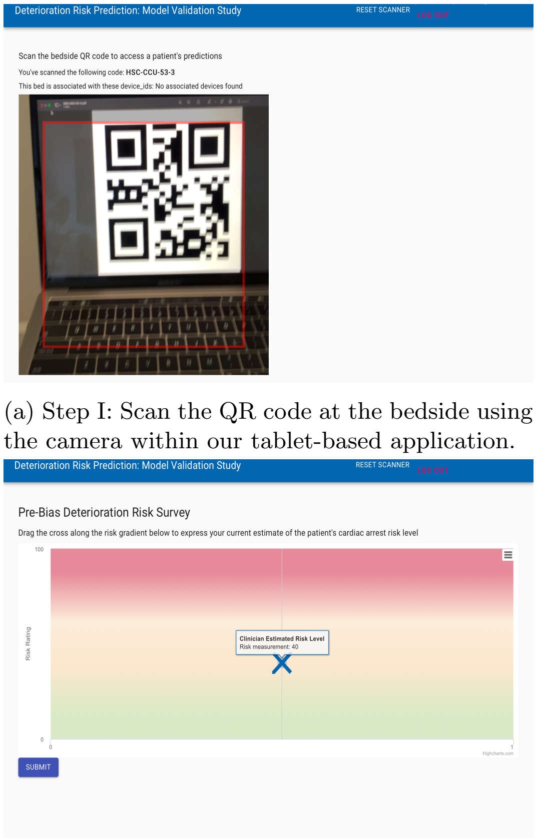
4.2. Implementing the protocol

The silent trial will span over a six-month period to gather enough samples to demonstrate the model's performance over time and give us a chance to assess the model on rare events. With around 200 unique monthly admissions to the ICU, a six-month trial will create opportunities to annotate risk for as many as 1200 unique patients. As part of the trial, the evaluators are responsible for assessing the risk estimates of the model. This is done through our user interface, and the steps are described in section 4.2.1

4.2.1. Validation Study Interface:

During the silent trial period, the evaluators use the user interface (UI) for validating the model and providing annotations. Figures 3a- 3e provide a breakdown of the validation workflow. Each bedspace in our study's ICU is labelled with a unique QR code that electronically maps the bed space to the patient that occupies that space in our database. In Step I, the evaluators will need to physically approach a given bedspace and scan its unique QR code, launching the patient-specific risk validation workflow. This step ensures evaluators are present at a patient's bedside and conducting their own clinical evaluation during the validation process. In Step II, evaluators are prompted to place the patient into a broad risk category, based on their clinical judgment. This step occurs each time a clinician scans a QR code, and is a necessary step to avoid the priming effect and anchoring bias brought on by observing the model's prediction. This estimate will be used in our analysis as the initial expert estimation of risk, and will also serve to estimate inter-evaluator risk baselines. This estimate will be compared to the evaluator's risk estimate after observing the prediction, allowing us to measure the possible impact of seeing a model's prediction on the perceived risk estimated of a clinician.

Upon entering their risk estimate, evaluators unmask the model's predicted risk trendline in Step III. We have chosen to display trends instead of just showing the latest risk estimate. This is because when considering the actionability of a model's visualization in healthcare settings, trend trumps individual prediction, and actionability is deemed inextricably linked to the perceived trustworthiness of the visualization (Harrigan and Morgenshtern et al., 2021). Evaluators are then prompted to validate the current risk estimate by dragging only the last estimated value to a risk category they consider more sensible. This evaluation will be compared to the risk estimate of the model to measure the model's performance over time. Finally, in the last step, the evaluators are encouraged to provide their rational for their adjusted risk by enter text annotations. This allows us to capture the reasons for which there may arise a disagreement between our model and the clinicians. Additionally, the evaluator may easily edit their current annotation. This allows for flexibility in how users may choose to complete an evaluation and provides us with insight into how such evaluations are performed within a clinician's workflow. Giving clinicians space to draft notes on their patient concerns prior to dragging the curve presents evaluators with an opportunity for cognitive offload. These annotations are also helpful for our analysis of failure modes. We will use the annotations to identify the categories of clinical complications that are better captured or often missed by the model. Our pipeline is also linked to the Electronic Health Records (EHR), allowing us to look into vulnerable patient populations and measure the performance across those individuals.


Deterioration Risk Prediction: Model Validation Study
RESET SCANNER
LOG OUT

Scan the bedside QR code to access a patient's predictions

You've scanned the following code: **HSC-CCU-53-3**

This bed is associated with these device_ids: No associated devices found

(a) Step I: Scan the QR code at the bedside using the camera within our tablet-based application.

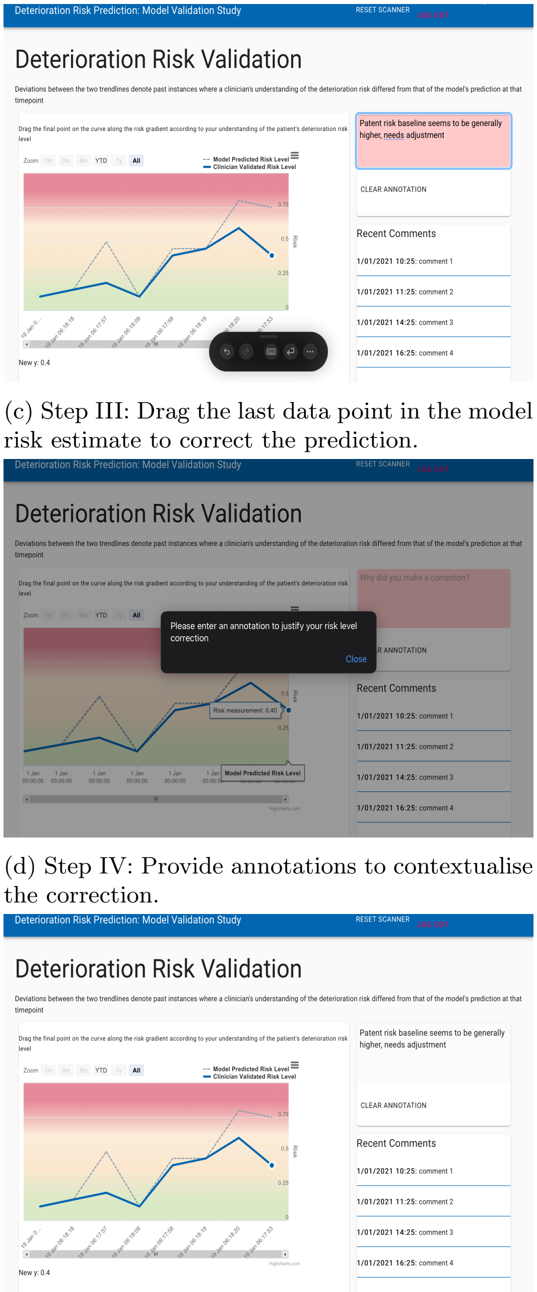
Deterioration Risk Prediction: Model Validation Study
RESET SCANNER
LOG OUT

Pre-Bias Deterioration Risk Survey

Drag the cross along the risk gradient below to express your current estimate of the patient's cardiac arrest risk level

(b) Step II: Drag the marker to an initial, pre-bias risk estimate.

Figure 3



Deterioration Risk Validation

Deviations between the two trendlines denote past instances where a clinician's understanding of the deterioration risk differed from that of the model's prediction at that timepoint

Drag the final point on the curve along the risk gradient according to your understanding of the patient's deterioration risk level

Patent risk baseline seems to be generally higher, needs adjustment

1/01/2021 10:25: comment 1

1/01/2021 11:25: comment 2

1/01/2021 14:25: comment 3

1/01/2021 16:25: comment 4

(c) Step III: Drag the last data point in the model risk estimate to correct the prediction.

Deterioration Risk Validation

Deviations between the two trendlines denote past instances where a clinician's understanding of the deterioration risk differed from that of the model's prediction at that timepoint

Drag the final point on the curve along the risk gradient according to your understanding of the patient's deterioration risk level

Why did you make a correction?

Please enter an annotation to justify your risk level correction

1/01/2021 10:25: comment 1

1/01/2021 11:25: comment 2

1/01/2021 14:25: comment 3

1/01/2021 16:25: comment 4

(d) Step IV: Provide annotations to contextualise the correction.

Deterioration Risk Validation

Deviations between the two trendlines denote past instances where a clinician's understanding of the deterioration risk differed from that of the model's prediction at that timepoint

Drag the final point on the curve along the risk gradient according to your understanding of the patient's deterioration risk level

Patent risk baseline seems to be generally higher, needs adjustment

1/01/2021 10:25: comment 1

1/01/2021 11:25: comment 2

1/01/2021 14:25: comment 3

1/01/2021 16:25: comment 4

(e) Users are able to edit their annotations, and read annotations made by other users for past datapoints.

Figure 3

4.2.2. Deployment Infrastructure

The real-time validation of our ML model requires it to be integrated into the hospital infrastructure. The lack of a systematic deployment solution for ML models in healthcare settings and the challenges in the integration of such models into the existing hospital infrastructures compels most researchers to design their own IT solution for deployment (Hong et al., 2020; Drysdale et al., 2020). For the purpose of our silent trial, we designed a scalable and customizable deployment pipeline for high-frequency physiological signals, as summarized in Figure 4. To the best of our knowledge, there are no widely followed systemic guidelines in healthcare at this point that enable easy integration for a variety of models. A typical deployment pipeline can be broken down into three key steps: 1) data collection, 2) data preparation, and 3) inference.

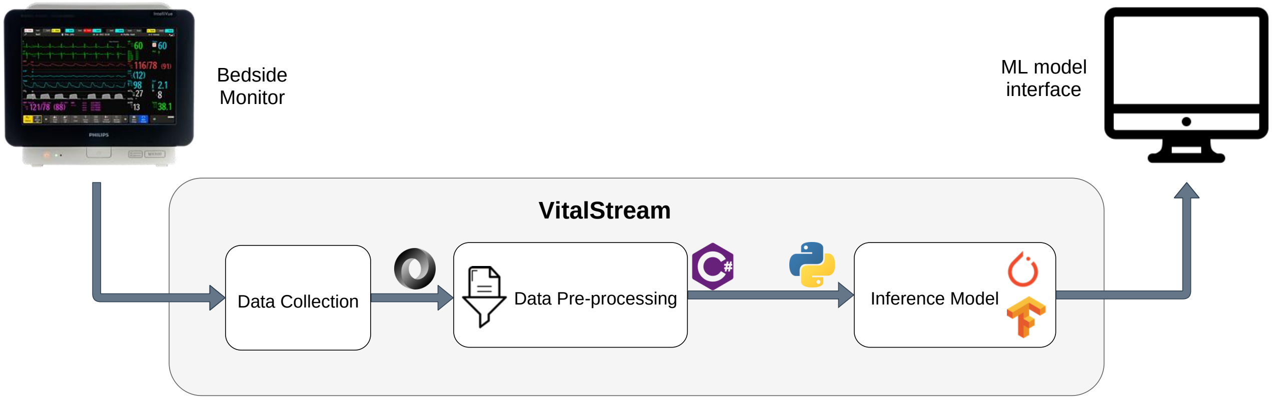


Figure 4: Deployment pipeline which includes 3 parts: 1) Data collection from bedside monitors, 2) data processing, 3) Inference model.

The first component is the **data collection/retrieval** system which provides access to retrospective data for offline training/testing and real-time data for online inference. Hospitals often use in-house data storage solutions for keeping retrospective data, following different standards. As for real-time measurements, bedside monitoring devices collect the data in different formats and communicate them with hospital servers using various protocols depending on the equipment manufacturer. Therefore, a unified data retrieval system should be capable of supporting a wide range of data formats and communication protocols. For this study, we retrieve multivariate physiological signals of the ICU patients through the existing data storage and retrieval system at the Hospital for Sick Children (HSC), Toronto, Ontario, Canada (Goodwin et al., 2020).

After retrieval, the data should be subjected to a series of data cleaning processes before being used for data analysis, since physiological data are riddled with noise, discontinuity, asynchronisation, and artifacts. Moreover, ML models might require additional feature extraction from the raw data (e.g. the systolic and diastolic blood pressure and heart rate measured from ECG) making **data pre-processing** an integral part of this pipeline. For models like ours, where near real-time prediction is essential, processing data in streams rather than batches is more suitable. There exist isolated solutions for components along the end-to-end pipeline. For instance, a number of existing approaches are proposed for addressing large scale stream processing demands (Van Der Veen et al., 2015b; Akidau et al., 2015; Ramadan, 2020) using popular stream processing engines such as Apache Storm (Van Der Veen et al., 2015a), Spark Streaming (Zaharia et al., 2016), and others. However, these engines are not specialized for physiological data processing and are $10-100\times$ slower than the numerical libraries such as SciPy, NumPy and Scikit-learn (Jayarajan et al., 2020). On the other hand, such numerical libraries lack the temporal query language support making it difficult for the researchers to write easy data pre-processing code. Considering the shortcomings in the existing solutions, we designed a pipeline named VitalStream, based on Microsoft Trill (Chandramouli et al., 2014), a state-of-the-art streaming engine specialized in processing high-volume, real-time streams. With the use of Trill, we wrote data processing queries that synchronizes data and handles artifacts (e.g. impute missing values, remove duplicates, etc). As some signals, such as heart rate, can be measured by more than one device at the same ICU bed, streams representing the same signal type are merged into a single stream during pre-processing. In addition, our pre-processing pipeline can perform various data processing steps, such as normalization, re-sampling, and real-time batching on input data windows. These data pre-processing steps are customizable and with the change of parameters can be used for other similar data streaming deployment projects.

As most of the ML models are built using Python-based machine learning frameworks, such as Tensorflow, and PyTorch, we serialize the pre-processed data streams using protocol buffers [1] and transfer

[1] https://developers.google.com/protocol-buffers

them to the Python environment via ZeroMQ [2] messaging library to be used for **inference** by the prediction model. The risk prediction generated by the model, as well as other identifying information (such as timestamp and device-ID), are stored in the MySQL [3] database for future retrieval and validation. The use of the publish-subscribe pattern in the communication between the pre-processing framework and the model also allows us to connect and evaluate multiple models on the same data, in real-time, without any additional data pre-processing or storage cost.

To test the performance of VitalStream, we measured its throughput by passing the streams of 11 signals from the messaging queues coming from the bedside monitors. On a single core, the pre-processing pipeline has a peak throughput of 3.9K input messages per second per signal stream. This is enough to handle more than a typical number of beds in an ICU and ensures that the pipeline will be able to run the inference model for all the patients at the same time. We also evaluated the latency, which measures the time it takes for messages to go through the pipeline. For this, we configured the pipeline to output 10-minute windows every 5 seconds using the 11 input streams of 1 Hz signals from the ICU. The 50th percentile latency is 4.61 seconds and the 90th percentile latency is 64.27 seconds, meaning VitalStream can run in real-time as measurements are collected.

5. Related Work

Despite the world class research in the field of ML in healthcare, only a small percentage of models are deployed in real-life hospital settings (Verma et al., 2021; Henry et al., 2017; Razavian et al., 2020; Sendak et al., 2020b; Connell et al., 2019; Levin et al., 2018; Kang et al., 2016; Giannini et al., 2019; Nestor et al., 2020). To validate performance and usability, all of these models have gone through rigorous real-time evaluation. Depending on the type of model and the clinical setting it is intended for, different validation studies have been designed and executed. Some run the models in real-time and in complete isolation, and evaluate their performance post trial. In such cases, they either have access to well-documented labels for what the model is predicting (Henry et al., 2017; Razavian et al., 2020; Brajer et al., 2020; Shamout et al., 2021), or use the model to identify the patient group of interest, and compare this list to what had been identified in the clinic post-trial (Kang et al., 2016; Parikh et al., 2019b; Bell et al., 2021).

Others take a more human-centred approach towards the silent trial and involve the clinicians in the validation. The importance of including users in the validation of clinical tools through participatory approaches is highlighted in H. Gyldenkaerne et al. (2020), where the data provided to train a tool predicting appointment no-shows was so cumbersome to enter into the EHR, that clinicians were unable to provide sufficiently detailed information for the model to make accurate predictions. To validate a tool identifying hospitalized patients at high risk for clinical deterioration, Escobar et al. (2020) conducted a multi-site validation study. This study shows that performing interventions based on the alarms triggered by the model resulted in a reduced mortality in hospitalized patients. They implemented an intervention program as a part of the trial where the nurses remotely monitored the condition of high risk patients and communicated those to rapid-response teams at the hospital. Similarly, for evaluating a sepsis prediction tool called Sepsis Watch, Sendak et al. (2020b) employs a Rapid Response team that monitors the model prediction from a *control room* and only escalates cases deemed high risk. These studies are excellent validations of usability of ML tools in clinical settings. However, they are not designed to evaluate the model in a silent mode, and they will impact the clinical workflow. In addition, none of these trials are designed with the goal of formative testing, nor do they collect information that can be used for design refinement.

Even though there has been a significant effort in outlining the general approach necessary for deployment and validation of clinical ML models (Lu et al., 2021; Antoniou and Mamdani, 2021), such efforts are often not backed with a practical use case. This gap may explain why only a subset of implemented models have been demonstrated to produce a meaningful clinical impact (Yin et al., 2021). The canonical pathway for any translational research (non-ML silent trials), rather, focuses on aligning evaluation with the information needs of the user. The silent trial approach we propose here is aligned with previous work describing the importance of preclinical studies, to better guide prospective evaluation (Campbell et al., 2007).

[2] https://zeromq.org
[3] https://www.mysql.com

6. Conclusion

Integration and translation of ML models into clinical practice is a complex process, and its success in organizational adoption relies on many contributors. This paper looks into the design and implementation of a silent trial that allows for rigorous evaluation of ML models in real-life hospital settings. Innovative trial designs of this sort are integral for assessing the safety and usability of ML tools prior to deployment in healthcare settings and will be required to validate models that provide estimates for clinical phenomena with no existing clinical labels. We believe that the framework for the trial we describe presents a scalable mechanism for both validating models and collecting information that assists in model refinement. Design and implementation of this framework would not have been possible without expertise and collaboration of multiple disciplines, including machine learning, medicine, human-computer interaction, human factors, and systems engineering. These allowed us to design a trial that thoroughly evaluates a complex clinical model from several relevant perspectives to our clinical setting. We hope that the principles introduced in this paper and the considerations that have gone into designing the silent trial protocol can help other researchers validate ML models in their clinical settings.

Institutional Review Board (IRB)

This work has an REB approval with reference number: 1000074229. No patients were included in this silent trial stage as the REB approved a waiver of consent given that there is (a) minimal risk given that no alterations to standard of care were made and (b) it would be impracticable to consent each patient given the lack of research resources in the ICU. For medical trainees and other research team members, agreement to participate in the validation components of this trial was voluntarily and recognized by compensations as approved by the REB. These activities are aligned with the goals of the institution, such as to facilitate research and quality improvement activities.

Acknowledgments

This work has been generously supported by the Canadian Institute of Health Research (CIHR) and the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

Tyler Akidau, Robert Bradshaw, Craig Chambers, Slava Chernyak, Rafael J Fernández-Moctezuma, Reuven Lax, Sam McVeety, Daniel Mills, Frances Perry, Eric Schmidt, et al. The dataflow model: a practical approach to balancing correctness, latency, and cost in massive-scale, unbounded, out-of-order data processing. 2015.

Tony Antoniou and Muhammad Mamdani. Evaluation of machine learning solutions in medicine. *CMAJ*, 193(36):E1425–E1429, 2021. ISSN 0820-3946. doi: 10.1503/cmaj.210036. URL https://www.cmaj.ca/content/193/36/E1425.

David Bell, John Baker, Chris Williams, and Levi Bassin. A trend-based early warning score can be implemented in a hospital electronic medical record to effectively predict inpatient deterioration. *Critical care medicine*, 49(10):e961, 2021.

Jacob N Blackwell, Jessica Keim-Malpass, Matthew T Clark, Rebecca L Kowalski, Salim N Najjar, Jamieson M Bourque, Douglas E Lake, and J Randall Moorman. Early detection of in-patient deterioration: one prediction model does not fit all. *Critical care explorations*, 2(5), 2020.

Nathan Brajer, Brian Cozzi, Michael Gao, Marshall Nichols, Mike Revoir, Suresh Balu, Joseph Futoma, Jonathan Bae, Noppon Setji, Adrian Hernandez, et al. Prospective and external evaluation of a machine learning model to predict in-hospital mortality of adults at time of admission. *JAMA network open*, 3(2):e1920733–e1920733, 2020.

John Brooke. Sus: a retrospective. *Journal of usability studies*, 8(2):29–40, 2013.

Neil C Campbell, Elizabeth Murray, Janet Darbyshire, Jon Emery, Andrew Farmer, Frances Griffiths, Bruce Guthrie, Helen Lester, Phil Wilson, and Ann Louise Kinmonth. Designing and evaluating complex interventions to improve health care. *Bmj*, 334(7591):455–459, 2007.

Badrish Chandramouli, Jonathan Goldstein, Mike Barnett, Robert DeLine, Danyel Fisher, John C. Platt, James F. Terwilliger, and John Wernsing. Trill: A high-performance incremental query processor for diverse analytics. *Proc. VLDB Endow.*, 8(4):401–412, December 2014. ISSN 2150-8097. doi: 10.14778/2735496.2735503. URL https://doi.org/10.14778/2735496.2735503.

Matthew M Churpek, Trevor C Yuen, Christopher Winslow, David O Meltzer, Michael W Kattan, and Dana P Edelson. Multicenter comparison of machine learning methods and conventional regression for predicting clinical deterioration on the wards. *Critical care medicine*, 44(2):368, 2016.

Alistair Connell, Hugh Montgomery, Peter Martin, Claire Nightingale, Omid Sadeghi-Alavijeh, Dominic King, Alan Karthikesalingam, Cian Hughes, Trevor Back, Kareem Ayoub, et al. Evaluation of a digitally-enabled care pathway for acute kidney injury management in hospital emergency admissions. *NPJ digital medicine*, 2(1):1–9, 2019.

Sharon E Davis, Thomas A Lasko, Guanhua Chen, Edward D Siew, and Michael E Matheny. Calibration drift in regression and machine learning models for acute kidney injury. *Journal of the American Medical Informatics Association*, 24(6):1052–1061, 2017.

E. Drysdale, E. Dolatabadi, C. Chivers, V. Liu, S. Saria, M. Sendak, J. Wiens, M. Brudno, A. Hoyt, M. Mazwi, M. Mamdani, D. Singh, V. Allen, C. McGregor, H. Ross, A. Szeto, A. A. Verma, B. Wangand P. A. Paprica, and A. Goldenberg. Implementing ai in healthcare. Technical report, Vector Institute, The Hospital for Sick Children, March 2020. URL https://vectorinstitute.ai/wp-content/uploads/2020/03/implementing-ai-in-healthcare.pdf.

Ely Erez, Mjaye L Mazwi, Alexandra M Marquez, Michael-Alice Moga, and Danny Eytan. Hemodynamic patterns before inhospital cardiac arrest in critically ill children: An exploratory study. *Critical care explorations*, 3(6), 2021.

Gabriel J Escobar, Vincent X Liu, Alejandro Schuler, Brian Lawson, John D Greene, and Patricia Kipnis. Automated identification of adults at risk for in-hospital clinical deterioration. *New England Journal of Medicine*, 383(20):1951–1960, 2020.

Milena A Gianfrancesco, Suzanne Tamang, Jinoos Yazdany, and Gabriela Schmajuk. Potential biases in machine learning algorithms using electronic health record data. *JAMA internal medicine*, 178 (11):1544–1547, 2018.

Heather M Giannini, Jennifer C Ginestra, Corey Chivers, Michael Draugelis, Asaf Hanish, William D Schweickert, Barry D Fuchs, Laurie Meadows, Michael Lynch, Patrick J Donnelly, et al. A machine learning algorithm to predict severe sepsis and septic shock: development, implementation, and impact on clinical practice. *Read Online: Critical Care Medicine— Society of Critical Care Medicine*, 47(11):1485–1492, 2019.

Andrew J Goodwin, Danny Eytan, Robert W Greer, Mjaye Mazwi, Anirudh Thommandram, Sebastian D Goodfellow, Azadeh Assadi, Anusha Jegatheeswaran, and Peter C Laussen. A practical approach to storage and retrieval of high-frequency physiological signals. *Physiological Measurement*, 41(3):035008, 2020.

Christopher H. Gyldenkaerne, Gustav From, Troels Mønsted, and Jesper Simonsen. Pd and the challenge of ai in health-care. In *Proceedings of the 16th Participatory Design Conference 2020-Participation (s) Otherwise-Volume 2*, pages 26–29, 2020.

Harrigan and Morgenshtern, Anna Goldenberg, and Fanny Chevalier. Considerations for visualizing uncertainty in clinical machine learning models. In *CHI '21 Workshop: Realizing AI in Healthcare: Challenges Appearing in the Wild*, 2021. URL http://francisconunes.me/RealizingAIinHealthcareWS/papers/Harrigan2021.pdf.

Katharine Henry, Shannon Wongvibulsin, Andong Zhan, Suchi Saria, and David Hager. Can septic shock be identified early? evaluating performance of a targeted real-time early warning score (trewscore) for septic shock in a community hospital: global and subpopulation performance. In *Critical Care: Do we have a crystal ball? predicting clinical deterioration and outcome in critically ill patients*, pages A7016–A7016. American Thoracic Society, 2017.

Katharine E Henry, David N Hager, Peter J Pronovost, and Suchi Saria. A targeted real-time early warning score (trewscore) for septic shock. *Science translational medicine*, 7(299):299ra122–299ra122, 2015.

Robert R. Hoffman, Matthew Johnson, Jeffrey M. Bradshaw, and Al Underbrink. Trust in automation. *IEEE Intelligent Systems*, 28(1):84–88, 2013. doi: 10.1109/MIS.2013.24.

Shenda Hong, Yanbo Xu, Alind Khare, Satria Priambada, Kevin Maher, Alaa Aljiffry, Jimeng Sun, and Alexey Tumanov. Holmes: Health online model ensemble serving for deep learning models in intensive care units. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1614–1624, 2020.

Anand Jayarajan, Kimberly Hau, Andrew Goodwin, and Gennady Pekhimenko. Lifestream: A high-performance stream processing engine for waveform data. *CoRR*, abs/2012.00192, 2020. URL https://arxiv.org/abs/2012.00192.

Dan W Joyce and John Geddes. When deploying predictive algorithms, are summary performance measures sufficient? *JAMA psychiatry*, 77(5):447–448, 2020.

Michael A Kang, Matthew M Churpek, Frank J Zadravecz, Richa Adhikari, Nicole M Twu, and Dana P Edelson. Real-time risk prediction on the wards: a feasibility study. *Critical care medicine*, 44(8):1468, 2016.

Arash Kia, Prem Timsina, Himanshu N Joshi, Eyal Klang, Rohit R Gupta, Robert M Freeman, David L Reich, Max S Tomlinson, Joel T Dudley, Roopa Kohli-Seth, et al. Mews++: enhancing the prediction of clinical deterioration in admitted patients through a machine learning model. *Journal of clinical medicine*, 9(2):343, 2020.

Scott Levin, Matthew Toerper, Eric Hamrock, Jeremiah S Hinson, Sean Barnes, Heather Gardner, Andrea Dugas, Bob Linton, Tom Kirsch, and Gabor Kelen. Machine-learning-based electronic triage more accurately differentiates patients with respect to clinical outcomes compared with the emergency severity index. *Annals of emergency medicine*, 71(5):565–574, 2018.

Charles Lu, Ken Chang, Praveer Singh, Stuart R. Pomerantz, Sean Doyle, Sujay Kakarmath, Christopher P. Bridge, and Jayashree Kalpathy-Cramer. Deploying clinical machine learning? consider the following.. *CoRR*, abs/2109.06919, 2021. URL https://arxiv.org/abs/2109.06919.

Melissa D McCradden, Shalmali Joshi, James A Anderson, Mjaye Mazwi, Anna Goldenberg, and Randi Zlotnik Shaul. Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning. *Journal of the American Medical Informatics Association*, 27(12):2024–2027, 2020.

Melissa D McCradden, James A Anderson, Elizabeth A. Stephenson, Erik Drysdale, Lauren Erdman, Anna Goldenberg, and Randi Zlotnik Shaul. A research ethics framework for the clinical translation of healthcare machine learning. *The American Journal of Bioethics*, pages 1–15, 2022.

Bret Nestor, Matthew BA McDermott, Willie Boag, Gabriela Berner, Tristan Naumann, Michael C Hughes, Anna Goldenberg, and Marzyeh Ghassemi. Feature robustness in non-stationary health records: caveats to deployable model performance in common clinical machine learning tasks. In *Machine Learning for Healthcare Conference*, pages 381–405. PMLR, 2019.

Bret Nestor, Liam G McCoy, Amol Verma, Chloe Pou-Prom, Joshua Murray, Sebnem Kuzulugil, David Dai, Muhammad Mamdani, Anna Goldenberg, and Marzyeh Ghassemi. Preparing a clinical support model for silent mode in general internal medicine. In *Machine Learning for Healthcare Conference*, pages 950–972. PMLR, 2020.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

Marcus Eng Hock Ong, Christina Hui Lee Ng, Ken Goh, Nan Liu, Zhi Xiong Koh, Nur Shahidah, Tong Tong Zhang, Stephanie Fook-Chong, and Zhiping Lin. Prediction of cardiac arrest in critically ill patients presenting to the emergency department using a machine learning score incorporating heart rate variability compared with the modified early warning score. *Critical Care*, 16(3): R108, 2012.

Ravi B Parikh, Stephanie Teeple, and Amol S Navathe. Addressing bias in artificial intelligence in health care. *Jama*, 322(24):2377–2378, 2019a.

Ravi Bharat Parikh, Chris Manz, Corey Chivers, Susan B Regli, Jennifer Braun, Joshua Adam Jones, Ronac Mamtani, Michael Draugelis, Justin E Bekelman, Amol S Navathe, et al. Derivation and implementation of a machine learning approach to prompt serious illness conversations among outpatients with cancer., 2019b.

Fawzya Ramadan. Real-time healthcare monitoring system using online machine learning and spark streaming. 01 2020.

Narges Razavian, Vincent J Major, Mukund Sudarshan, Jesse Burk-Rafel, Peter Stella, Hardev Randhawa, Seda Bilaloglu, Ji Chen, Vuthy Nguy, Walter Wang, et al. A validated, real-time prediction model for favorable outcomes in hospitalized covid-19 patients. *NPJ digital medicine*, 3(1):1–13, 2020.

Jeff Rubin and Dana Chisnell. How to plan, design, and conduct effective tests. *Handbook of usability testing*, 17(2):348, 2008.

Jefferey Rubin and D Chisnell. Usability testing: How to plan, design and conduct effective tests, 1994.

Craig G Rusin, Sebastian I Acosta, Lara S Shekerdemian, Eric L Vu, Aarti C Bavare, Risa B Myers, Lance W Patterson, Ken M Brady, and Daniel J Penny. Prediction of imminent, severe deterioration of children with parallel circulations using real-time processing of physiologic data. *The Journal of thoracic and cardiovascular surgery*, 152 (1):171–177, 2016.

Christopher B Sankey, Gail McAvay, Jonathan M Siner, Carol L Barsky, and Sarwat I Chaudhry. “deterioration to door time”: an exploratory analysis of delays in escalation of care for hospitalized patients. *Journal of general internal medicine*, 31 (8):895–900, 2016.

Mark P Sendak, Joshua D’Arcy, Sehj Kashyap, Michael Gao, Marshall Nichols, Kristin Corey, William Ratliff, and Suresh Balu. A path for translation of machine learning products into healthcare delivery. *European Medical Journal Innovations*, 2020a.

Mark P Sendak, William Ratliff, Dina Sarro, Elizabeth Alderton, Joseph Futoma, Michael Gao, Marshall Nichols, Mike Revoir, Faraz Yashar, Corinne Miller, et al. Real-world integration of a sepsis deep learning technology into routine clinical care: implementation study. *JMIR medical informatics*, 8 (7):e15182, 2020b.

Farah E Shamout, Yiqiu Shen, Nan Wu, Aakash Kaku, Jungkyu Park, Taro Makino, Stanislaw Jastrzebski, Jan Witowski, Duo Wang, Ben Zhang, et al. An artificial intelligence system for predicting the deterioration of covid-19 patients in the emergency department. *NPJ digital medicine*, 4 (1):1–11, 2021.

Clay Spinuzzi. The methodology of participatory design. *Technical communication*, 52(2):163–174, 2005.

Adarsh Subbaswamy and Suchi Saria. From development to deployment: dataset shift, causality, and shift-stable models in health ai. *Biostatistics*, 21 (2):345–352, 2020.

Sana Tonekaboni, Mjaye Mazwi, Peter Laussen, Danny Eytan, Robert Greer, Sebastian D Goodfellow, Andrew Goodwin, Michael Brudno, and Anna Goldenberg. Prediction of cardiac arrest from physiological signals in the pediatric icu. In *Machine Learning for Healthcare Conference*, pages 534–550, 2018.

Jan Sipke Van Der Veen, Bram Van Der Waaij, Elena Lazovik, Wilco Wijbrandi, and Robert J. Meijer. Dynamically scaling apache storm for the analysis of streaming data. In *2015 IEEE First International Conference on Big Data Computing Service and Applications*, pages 154–161, 2015a. doi: 10.1109/BigDataService.2015.56.

Jan Sipke Van Der Veen, Bram van der Waaij, Elena Lazovik, Wilco Wijbrandi, and Robert J Meijer. Dynamically scaling apache storm for the analysis of streaming data. In *2015 IEEE First International Conference on Big Data Computing Service and Applications*, pages 154–161. IEEE, 2015b.

Amol A Verma, Joshua Murray, Russell Greiner, Joseph Paul Cohen, Kaveh G Shojania, Marzyeh Ghassemi, Sharon E Straus, Chloe Pou-Prom, and Muhammad Mamdani. Implementing machine learning in medicine. *CMAJ*, 193(34):E1351–E1357, 2021.

Ben Wellner, Joan Grand, Elizabeth Canzone, Matt Coarr, Patrick W Brady, Jeffrey Simmons, Eric Kirkendall, Nathan Dean, Monica Kleinman, and Peter Sylvester. Predicting unplanned transfers to the intensive care unit: a machine learning approach leveraging diverse clinical elements. *JMIR medical informatics*, 5(4):e8680, 2017.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. Do no harm: a roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9):1337–1340, 2019.

Qian Yang, Aaron Steinfeld, and John Zimmerman. 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*, pages 1–11, 2019.

Jiamin Yin, Kee Yuan Ngiam, and Hock Hai Teo. Role of artificial intelligence applications in real-life clinical practice: Systematic review, Apr 2021. URL https://www.jmir.org/2021/4/e25759.

Matei Zaharia, Reynold S. Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J. Franklin, Ali Ghodsi, Joseph Gonzalez, Scott Shenker, and Ion Stoica. Apache spark: A unified engine for big data processing. *Commun. ACM*, 59(11):56–65, oct 2016. ISSN 0001-0782. doi: 10.1145/2934664. URL https://doi.org/10.1145/2934664.

Appendix A. Appendix

A.1. Risk estimator model

The risk estimator is a deep learning model, composed of a convolutional network trained end-to-end with a recurrent layers. The architecture is based on the model introduced in Tonekaboni et al. (2018), with a few modifications. It estimates a risk probability over time using a 10 minute history of signals collected from the bedside monitors. Table 1 provides the list of signals used and details about the measurement frequency. Unlike the original model, this is not an ensemble and each prediction is made based on the 10 input physiological signals in aggregate. The model updates its prediction every minute as new measurements come in. Since missing measurements are very common in healthcare data, the model takes in an additional mask channel, indicating which measurements were taken and which ones are missing and therefore were imputed. The model takes in this information and incorporates it into its prediction. We use Cardio Pulmonary Arrest (CPA) labels, as proxies for high risk events, in order to train the model.

Signal	Description	Sample Rate
HR	Heart rate derived from ECG	5 seconds
RR	Resp Respiratory rate	5 seconds
Pulse	Pulse rate	5 seconds
SPO2	Oxygen saturation level	5 seconds
etCO2	End-tidal CO2	5 seconds
CVPm	Central Venus Pressure	5 seconds
AWRR	airway respiratory rate	5 seconds
NBPm	Non-invasive blood pressure	1-60 minutes
NBPd	Non-invasive blood pressure	1-60 minutes
NBPs	Non-invasive blood pressure	1-60 minutes

Table 1: Description of high-frequency ICU signals

A.2. Usability Testing

The usability study itself is comprised of 2 phases: usability testing of the general workflow, and a focus group follow-up discussion on clinicians' understanding of the various terms and concepts in the interface. The first aims to compare pre-bias risk survey response to users' interactions on the final prediction trend page under several clinical contexts, and the second aims to learn whether the presented approach is realistic in the introduction of clinical ML. Prior to participating in the focus group, each of the participants in the segment completes an augmented System Usability Scale (SUS) questionnaire (Brooke, 2013). This collects for us some standardised measure of the appeal of the workflow with regards to its ease-of-use; augmentation includes several additional questions appended to the end of the published SUS that are specific to additional interface details for we desire user feedback. The focus group chiefly works to determine whether, after using the model on several clinical scenarios of variable risk trend, users develop sufficient trust in the validation interface to deem the silent trial effort trustworthy. There must be adequate trust for the adoption of the system and for robust stakeholder engagement in the study (Hoffman et al., 2013). The group setting of the discussion hopes to achieve enthusiastic discourse on this topic between the clinicians, in a comfortable environment for fostering a lively debate amongst the segment participants.