The HEART Interface: Visualizing Risk Score Uncertainty in the Cardiothoracic ICU

Mahsan Nourani
The Roux Institute
Northeastern University
Portland, Maine, USA
m.nourani@northeastern.edu

Lien Nguyen
Khoury College
Northeastern University
Boston, Massachusetts, USA
nguyen.lien@northeastern.edu

Carey Barry
Northeastern University
Boston, Massachusetts, USA
c.barry@northeastern.edu

Qingchu Jin
Roux Insititute
Northeastern University
Portland, Maine, USA
q.jin@northeastern.edu

Melanie Tory
Roux Institute
Northeastern University
Portland, Maine, USA
m.tory@northeastern.edu

Abstract

Artificial Intelligence (AI) holds significant potential for supporting clinical decision-making, particularly in high-pressure environments, such as Cardiothoracic Intensive Care Units (CT-ICU). Care teams in these settings face challenges such as alarm fatigue, rapid staff turnover, time-sensitive decisions, and an overwhelming amount of data. AI-driven Clinical Decision-Support Systems (AI-CDSS) can support care teams in overcoming some of these challenges by providing solutions like detecting and reporting risk scores for adverse events that may lead to increased fatalities or re-admissions, enabling timely intervention. One key challenge with risk scores is missing data, which can create considerable uncertainty in risk score values. AI-CDSSs rarely convey the risk score uncertainty, which is important in the effectiveness and reliability of clinical decision-making. In this paper, we describe the interface design process for *HEART*, an AI-powered system developed collaboratively with clinical and AI experts over a 16-month iterative design process for a hospital's CT-ICU. The *HEART* interactive interface integrates understandable visualizations of risk scores and their uncertainty within both a holistic view of all patients in the unit and detailed patient-specific views. We reflect on the user-centered design process, report findings from an expert walkthrough study, and discuss lessons learned as well as broader implications. This work contributes valuable insights into uncertainty visualization design for AI-derived risk scores in a critical care application. Beyond these specific insights, our work illustrates the kind of comprehensive, human-centered design process necessary for responsible AI adoption in critical environments.
Supplemental Material, including a video demo of the *HEART* interface and additional details on the algorithm, is available on the project's OSF repository: https://osf.io/akqx8/.

IUI '26, Paphos, Cyprus

ACM ISBN 979-8-4007-1984-4/26/03
https://doi.org/10.1145/3742413.3789109

CCS Concepts

• **Human-centered computing** → **Interactive systems and tools**; **HCI design and evaluation methods**; **Visual analytics**; **User centered design**; *Participatory design*; **Interface design prototyping**; • **Computing methodologies** → *Artificial intelligence*; • **Applied computing** → **Health informatics**.

Keywords

AI in Healthcare, User-centered Design, AI-Supported Clinical Decision Support Systems, Human-centered AI, Explainable AI

ACM Reference Format:
Mahsan Nourani, Lien Nguyen, Carey Barry, Qingchu Jin, and Melanie Tory. 2026. The HEART Interface: Visualizing Risk Score Uncertainty in the Cardiothoracic ICU. In *31st International Conference on Intelligent User Interfaces (IUI '26), March 23–26, 2026, Paphos, Cyprus.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3742413.3789109

1 Introduction

Artificial Intelligence (AI)-driven Clinical Decision Support Systems (AI-CDSS) enhance healthcare providers' ability to deliver patient care by identifying patterns in patient data that might otherwise go unnoticed by care teams [18]. AI-CDSS can help ensure more people receive preventive care, reduce delays in diagnosis, and promote a more consistent use of treatment guidelines [8]. In cardiovascular care, where quick intervention is essential, AI-CDSS can assist medical professionals in risk assessment, diagnosis refinement, and patient monitoring, resulting in more individualized and effective treatment approaches [8].

In this paper, we describe the interface design and evaluation of *HEART*: an interactive, visual tool (see Figure 1) for AI-supported monitoring of patients who are admitted to Cardiothoracic Intensive Care Units (CT-ICUs) following cardiac surgery. *HEART* was designed to support care teams in the CT-ICU by generating risk scores for adverse patient outcomes, enabling early detection and intervention. Its design is the culmination of a 16 month collaborative design process, involving Human-Computer Interaction (HCI) experts, health AI experts, a healthcare technology company, and a CT-ICU care team.

Failing to center HCI and user centered design in the development of AI-CDSSs has been frequently cited as a reason why they

often struggle with adoption in clinical practice [11, 27, 34, 35, 60]. As a result, there has recently been considerable attention in the HCI community towards designing for healthcare applications, including critical care medicine [60]. There is strong emphasis in the literature that AI should be treated as just one component of a complex sociotechnical system [31, 51] and on the importance of participatory design approaches that ensure AI products ultimately meet stakeholder needs [11, 60]. While AI can be employed for many different purposes in critical care, including resource management, diagnosis, and treatment planning, we focus systems that predict adverse events, reflected in AI-generated risk scores.

One key challenge in communicating AI predictions is to also convey their uncertainty, especially due to missing data. Risk score prediction relies heavily on the quality and completeness of the data. Unfortunately, missing data is a common issue, often due to clinical workflows [26] such as delayed or outdated test results, unavailable historical data from previous care settings (e.g., another hospital), or data that is observed locally but never entered into the system or chart, particularly with the increasing charting burden on care teams. This scarcity of data, driven by numerous factors, leads to uncertainty in predicted risk scores. Communicating the uncertainty information is critical to enabling care teams to calibrate their trust in an AI prediction and assess whether it should influence their care decisions. Yet how to effectively represent many risk scores, each with uncertainty, in a concise visual form, is an open research problem.

Existing design approaches for risk score visualization typically (1) focus on representing a single adverse event (e.g., sepsis) instead of multiple potential outcomes and/or (2) provide limited or no communication of risk score uncertainty [36]. ***HEART* is designed to address both of these gaps within a single system. It enables care teams to monitor all patients in the unit while offering individual views for each patient, tracks multiple adverse outcomes simultaneously, and develops techniques to communicate explanatory information and uncertainties due to missing values for all adverse events across all the patients.** The *HEART* interface is the result of a 16-month design study by a design team comprising HCI experts, AI experts, and a Physician Assistant with experience in cardiac post-surgical care, in close collaboration with a CT-ICU care team. Key contributions of our work are:

- Demonstration of a comprehensive design process for a clinical AI interface, involving close and prolonged collaboration with a multidisciplinary team of care professionals.
- Design of a visual interactive interface that conveys risk scores of many predicted outcomes for multiple monitored patients. The design comprises two main views: an aggregate view of adverse events across all patients and a patient-specific detail view.
- Exploration and evaluation of visual encodings for conveying uncertainty in risk scores and how the uncertainty interacts with categorized risk levels.
- Insights from an expert walkthrough study with CT-ICU care teams across various roles.

2 Related Work

We first describe the CT-ICU context to situate our design study. We then summarize key research on designing for critical care, designing for trust and reliance in AI within this domain, and visualization approaches for representing clinical risk scores.

2.1 CT-ICU Context

Survival outcomes of cardiac surgery patients are significantly influenced by postoperative complications [30, 41], including (with incidence rates): acute kidney injury (16%), renal failure (2%), new postoperative atrial fibrillation (32%), prolonged ventilation (10%), reoperation (6%), operative mortality (3%), stroke (ranging from 0.8% to 5.2% depending on the procedure), delirium (26% to 52%), and hospital stays longer than 14 days (8%) [19, 37, 41, 54]. Experiencing any of these adverse outcomes can cause death and morbidity, prolonged hospitalization, and increased hospital costs [41]. AI risk score predictions can support early intervention to reduce or mitigate these negative outcomes.

CT-ICU care teams are comprised of professionals with diverse backgrounds, roles, and expertise. Members range from bedside nurses needing immediate patient monitoring data to specialists requiring detailed longitudinal analyses. They are often overworked, alarm-fatigued, and need to make fast and accurate decisions while caring simultaneously for multiple patients. Care teams face significant cognitive load challenges, with studies showing that 80% of decision-making errors stem from information overload [20]. Turnover is also common in CT-ICUs, resulting in a frequent influx of new staff, especially when traveling nurses are brought in to fill staffing gaps. Statistical and technical literacy can vary considerably [12]. Some staff members may be well versed with interpreting statistics and visualizations but the majority lack this expertise; the cognitive effort required to understand complex statistical data and visual representations can be overwhelming. This mental load detracts from the primary focus of providing patient care. As such, developing easily understandable interfaces to convey AI-based health predictions, and the reasoning behind them, is a critical open research problem.

2.2 Designing for Critical Care

Designing for critical care medicine has unique challenges. Studies show 80% of "user errors" in clinical settings stem from cognitive overload [20], a particularly acute problem in the CT-ICU. Clinicians face constant challenges from false alarms, high "on-call" demands leading to sleep deprivation, and the need to make rapid decisions under time pressure [2, 34]. Given these workload demands, Yang et al. [59] recommend "unremarkable" AI design that fits into user routines in an unobtrusive way. Klüber et al. [35] similarly emphasize the need for AI-CDSS to meet a real user need and fit into workflows.

Prior work suggests that effective AI-CDSS should be designed to support all members of the care team, across various roles, domain expertise, and specific needs in delivering patient care [9, 12, 34] as well as collaboration across various roles. However, differences in roles and experience can affect users' trust formation and calibration, understanding, and decision-making behavior [45, 46]. Additionally, the various clinical roles may have different data tasks

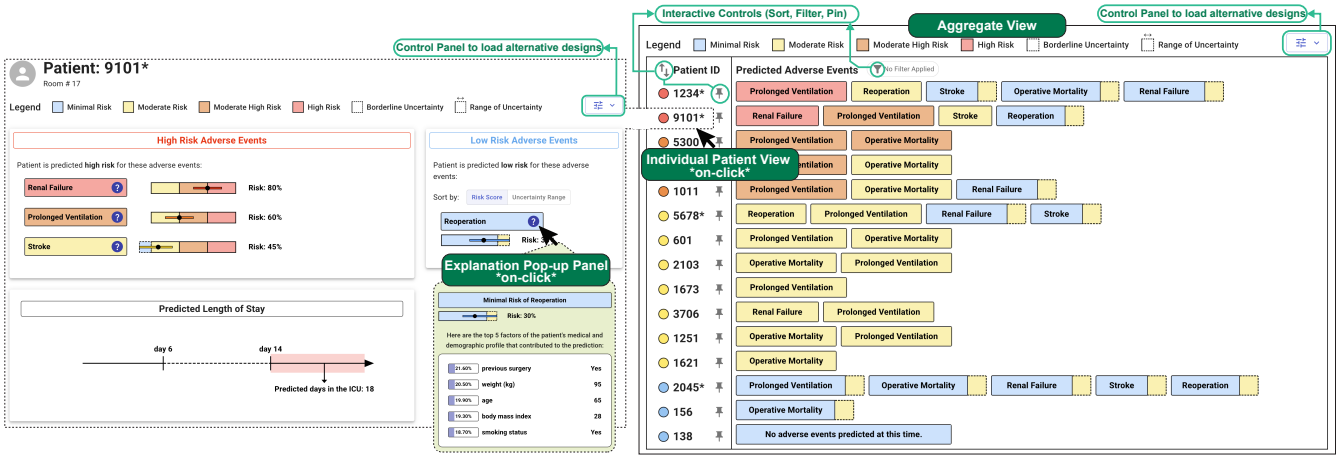


Figure 1: The main interface outlines for *HEART*. The interface comprises two main views: an Aggregate View (Right) offering an overview of all CT-ICU patients and their at-risk adverse events, and an Individual Patient View (Left) delivering detailed insights, including novel visualizations for risk and uncertainty, predicted length of stay, and feature-based (SHAP) explanations for each adverse event that are available on-demand. The interactive interface enables care teams to customize the system to suit their specific needs. We designed and tested alternative versions of certain interface elements, such as length of stay predictions and uncertainty boundaries, which can be adjusted through the control panel at the top right of each view. These alternatives are discussed in detail and illustrated in various figures throughout the paper.

and information needs that would benefit from bespoke designs or adaptation [34].

In a participatory design ideation study for the ICU, Yildirim et al. [60] reported numerous applications that would benefit from predictive models, including logistics support, dynamic staffing, classification of alerts according to urgency, and order recommendations. They also reported that physicians wanted to see the rationale behind AI predictions, a finding echoed by others [11, 34]. For instance, Cai et al. [11] reported that clinicians need holistic insight into an AI's capabilities, biases, and limitations. Based on a field study in the ICU, Yildirim et al. [61] reported that an AI for predicting when patients can be removed from ventilation would be most useful if integrated with the electronic medical record and if both patient-detail and unit-level overviews were available.

2.3 AI Trust, Reliance and Adoption

The high stakes in clinical decision-making coupled with errors that may have real-world and fatal consequences highlight the need for AI decision-making transparency and outcome explainability. Prior work has shown that more transparent algorithms can help users better understand how AI systems operate [10, 43], though the approach to transparency must be carefully balanced as increased transparency does not always lead to better decision-making outcomes [48]. Moreover, the quantity and depth of explanations [10, 48] and how they are presented [44, 45] affect how much users form trust and reliance in the AI outputs. Sivaraman et al. [53] report four observed interactions patterns for clinicians using AI treatment recommendations: ignore, negotiate, consider, and rely. They found that "negotiate"—which refers to when a medical decision-maker weighs and prioritizes individual aspects of the recommendations to follow or adjust—is the most appropriate form of trust calibration.

While most approaches to trust calibration focus on the AI providing explanation for its recommendations, Yang et al. [58] focused instead on having the AI provide supporting and opposing evidence from the literature. Other researchers are critical of XAI approaches in general. Ghassemi et al. [21] claim that XAI creates a "veneer of authenticity"; they focus instead on rigorous model validation. Miller et al. [42] propose "evaluative AI", an approach that engages with the user to provide evidence for and against an idea, but which never provides a direct recommendation.

Social factors and clear clinical value are also important factors influencing trust and adoption of AI in clinical contexts. Henry et al. [28] conducted interviews with clinicians who had used an AI system for sepsis for 6 months. They reported that physicians saw the AI as playing a supporting role in decision making, and that they built trust in the model by relating the training population to their patient population, through endorsement by a colleague or department head, and by observing model behavior in various situations. Similarly, Sendak et al.'s [51] deployment study with *SepsisWatch* showed how nursing staff calibrated their trust in the model over time and developed practices for working with the model. Based on their experience, Sendak et al. emphasize the importance of building relationships with stakeholders and including them throughout design, as well as treating the AI model as a decision aid rather than a replacement for human judgment. Zając et al. [62] presented design recommendations to support AI adoption for radiology, including enabling users to adjust functionality parameters, AI thresholds, and XAI methods. They also noted that provider workload and expertise should influence what information an AI provides. Similarly, Cai et al. [11] documented user onboarding needs for AI in a

pathology context and Jacobs et al. [31] emphasized the importance of designing AI as part of a multi-user collaborative system.

2.4 Visualization of Risk Scores and Uncertainty

Multiple systems have employed basic timeline visualizations (line, dot, and/or Gantt charts) [5, 13, 39, 47, 57] or heatmaps [53] of patient metrics, events, and a risk prediction score. Similar to our work, Ayobi et al. [3] visualize a risk value relative to low, medium, and high risk categories using a mark overlaid on a segmented bar chart. Some of these tools also include Shapley values to explain feature importance in relation to the risk score [3, 5, 13, 53, 57]. For instance, in the cardiac domain, *CardioAI* [57] introduced a dashboard to present a cardiotoxicity risk score timeline along with feature importance values and associated wearable sensor data. Bhattacharya et al. [7] developed an interactive dashboard for diabetes risk prediction that incorporated a variety of visual displays and counterfactual explanations. A study with caregivers found that color-coded risk factors and distribution charts were particularly helpful in identifying actions that could reduce diabetes risk. Preference studies indicate that the best visual encoding may vary depending on user roles, tasks, and clinical presentation, such as in Chiang et al.'s [15] survey of care providers and patients on visualization design preferences for seizure forecasting. All of these prior systems provide a useful starting point for our work; however, they showed prediction values for only one clinical outcome and one patient at a time and did not present any uncertainty information based on missing data.

Closer to our own work are approaches to visually encoding uncertainty information along with risk scores. While risk scores themselves represent a probability, there is also additional uncertainty in their calculation, which may be caused by noisy data, systematic model uncertainties (e.g., errors), and missing data (i.e., missing data critical in decision-making from patients) [26]. With regard to missing data, the magnitude of the uncertainty can be quantified and visualized. Most relevant to our work is Zhang et al.'s [63] *SepsisLab*, a system that calculated sepsis risk scores along with uncertainty due to missing data. Risk scores were visually encoded as a line chart and uncertainty was encoded as colored band surrounding the line. The system also recommended lab tests that would have the greatest impact on narrowing the risk prediction uncertainty band. A similar uncertainty band design was used by Yang et al. [59] to visualize uncertainty in survival rate predictions.

Conveying uncertainty could also be achieved through visualizations that communicate a range or distribution of predictions, such as error bars or violin plots [64]. Additionally, interactive and natural language approaches, such as explicit warnings and bias disclosures, can be employed to alert users to potential uncertainties or limitations in the model's outputs [24].

Our work on the *HEART* interface integrates new visualization approaches to convey prediction uncertainty, in relation to threshold risk levels, in a simple and concise manner. These concise representations open the possibility of conveying uncertainty information within a multi-patient overview and for multiple predicted outcomes. Thus, *HEART* integrates both a patient-level detail view and a unit-level overview as advocated by prior research in clinical settings [38, 61].

3 Design Study Process

The interface is one outcome stemming from a longstanding collaboration between academic HCI and AI experts (the authors of this paper), a CT-ICU (at *MaineHealth Maine Medical Center*[1]), and a digital and medical health software company (*Nihon Kohden*[2]). The long term goal of the collaboration is to develop novel algorithms and interfaces that accurately predict and explain negative patient outcomes from live patient data, deploy those tools in the CT-ICU, and validate their effectiveness through a clinical. In this paper, we focus on the design of the *HEART* interface, and in particular, the integration of risk score uncertainty (*For additional information on the broader HEART project, see Mazhude et al. [41], and see Jin et al. [32] for details on the HEART's predictive algorithm*).

Evidence indicates that involving healthcare professionals in the design of predictive technologies from the beginning can facilitate adoption and improve effectiveness [49]. To ensure the target users—CT-ICU care teams comprised of various roles and expertise—could fully benefit from the tool and to meet the design goals, we adopted a participatory, human-centered AI design approach. We collaborated closely with clinical and AI colleagues, through a design study process [50], over a period of 16 months. Our design process began with a 3 day intensive field study at the CT-ICU [6], consisting of contextual observation of key CT-ICU activities and interviews with 14 end users across various roles (bedside nurses, charge nurses, advance practice practitioners (APPs), intensivists, and surgeons). Findings from the field study identified key user roles and tasks, and informed the design goals outlined in Section 4.

We then conducted an iterative design process with frequent input from our clinical stakeholders, beginning with static wireframes and eventually implementing an interactive working prototype. Throughout the project, the research team conducted regular sessions with diverse care team members from the CT-ICU at *MaineHealth Maine Medical Center* to gather their feedback rapidly. To maximize participation while minimizing disruption, we scheduled these sessions during weekday lunch breaks on days when our contact charge nurse indicated lower-than-usual workload. Sessions took place in the CT-ICU break room, which featured a large display for demonstrations. We provided printed color versions of interface design alternatives and encouraged participants to annotate them directly. To accommodate the care team's demanding schedules, we maintained an informal, flexible structure—allowing participants to join and leave as needed, particularly when emergencies arose. We complemented these in-person sessions with virtual focus groups via Zoom for interested clinicians who had more availability. These remote sessions enabled more focused one-on-one discussions. We also accepted asynchronous feedback via email from clinicians who preferred time to reflect on the designs. This flexible, multi-modal approach ensured we gathered comprehensive feedback while respecting the clinical team's time constraints and unpredictable schedules.

We did not predetermine the number or frequency of feedback sessions. Instead, we followed an iterative approach: the team worked on design and implementation internally, addressing feedback from previous sessions, developing new features, and resolving

[1] https://www.mainehealth.org/maine-medical-center
[2] https://www.nihonkohden.com/

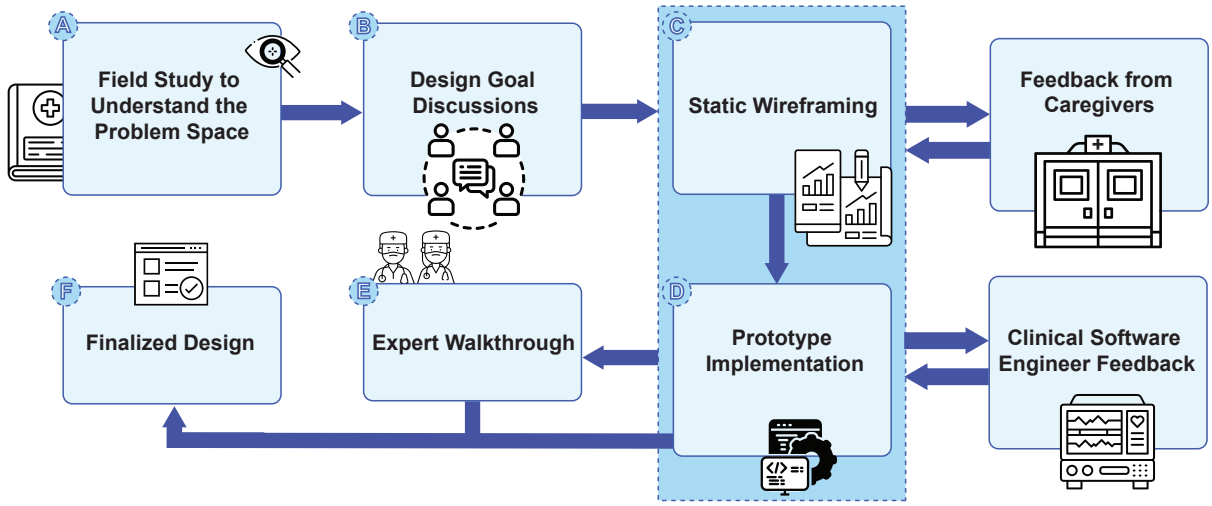


Figure 2: The interface design process for *HEART* involved iterative collaboration over 16 months. (A) We reviewed findings from our team's prior field study [6] and visited the *MaineHealth Maine Medical Center*'s CT-ICU to understand care team workflows and define design goals. (B) Our research team (including HCAI, AI, and clinical experts) established core design goals that were used as the basis for static wireframing. (C) During wireframing, we frequently visited the CT-ICU and sought feedback from the care team, and held weekly internal meetings to discuss and brainstorm ideas. (D) After several months of iterative refinement, we transitioned to implementing the prototype to identify and address flaws, as many designs depended on real data. During this phase, we held additional biweekly discussions with a clinical software company to gain feedback on design assumptions, optimize ML data formatting, and improve interface-to-backend integration. In addition to the CT-ICU care team, we received insights from two company employees with ICU experience, who contributed valuable perspectives on development complexities and care challenges. These individuals later participated in our expert walkthrough study. (E) Expert walkthroughs with CT-ICU five clinicians assessed the semi-final design, leading to final refinements and adjustments. These updates conclude our design process, with the finalized prototype now being integrated with live data and AI systems (F).

usability issues. We scheduled new sessions when we had stable updates ready for feedback or when we reached design decision points requiring additional input from clinicians. This iterative process naturally surfaced and resolved numerous usability issues throughout development. For example, we identified and simplified confusing interface interactions and redesigned how certain clinical outcomes were organized and presented. The usability discussions also generated several design alternatives, all described in Section 5. After collecting each round of feedback, the team then discussed the feedback internally and identified ways to integrate it into the system. Additionally, one member of our design team has a clinical background and experience working in similar intensive care settings. This iterative process, carried out over several months, enabled continuous incremental improvements to the interface, ensuring it aligned with the care team's needs and workflows. Figure 2 illustrates the design process, providing additional details for each step.

Naturally, after several iterative feedback cycles, the amount of new feedback decreased, indicating that we had reached feedback saturation in the design process. At this stage, we conducted an expert walkthrough study with target stakeholders. The discussions during the walkthrough study led to some updates and minor improvements to the interface design, also shown to the participants available for a quick follow-up feedback session. In Section 4, we present our design goals for the interface. Then, in Section 5, we describe the implemented system, which incorporates insights from informal expert consultations. We then present a formal expert walkthrough study in Section 6 to evaluate the system.

4 Design Goals

Based on our our understanding of the design problem and context arising from our field study [6] and related work, we identified *five* major design goals for our proposed user interface, which we will describe in this section.

Design Goal 1: Present Information and Predictions Simply for Rapid Comprehension. In the CT-ICU, care teams are often faced with time pressure and large volumes of data [34]; cognitive load is a substantial problem leading to decision-making errors [20]. To address this, an effective AI-CDSS must deliver information quickly and clearly, enabling the timely identification, interpretation, and response to adverse events in fast-paced clinical settings. Our system will emphasize visual presentation of outputs, insights, and data in a way that avoids overwhelming users. Additionally, recognizing that not all users are proficient in interpreting complex statistics and visualizations [12], the design will prioritize accessibility and ease of understanding for a diverse range of users.

Design Goal 2: Facilitate Monitoring of Various Adverse Events and Their Risk. Our design should focus on monitoring

and detecting key adverse events that are common causes of mortality and morbidity. We aim to incorporate the most frequent events identified by the ACSD Operative Risk Calculator from the Society of Thoracic Surgeons (STS), widely used by cardiac surgeons [55]. These include events such as *stroke*, *renal failure*, and *prolonged ventilation*. The interface will need to display the model-predicted risk of each adverse event for each patient, allowing care teams to quickly assess potential issues and take timely action.

Design Goal 3: Provide Comprehensive Patient Monitoring with On-Demand Details. Our goal is to design an AI-CDSS that enables care teams in CT-ICU units to efficiently monitor adversities among *all* patients while managing multiple cases simultaneously. The proposed interface will provide a clear overview of the entire patient population, with the ability to access detailed information on individual patients at the bedside or as needed, as recommended in prior work [38, 61]. To minimize information overload, less critical or specific details will be hidden by default and made available on demand following visualization design best practices.

Design Goal 4: Align With care teams' Existing Mental Models and Practices. Because CT-ICU care teams are often overloaded with information [20, 34], we aimed to minimize information burden and simplify learnability, by aligning with their existing terminology and mental models where possible. This aligns with Yang et al.'s notion of "unremarkable" AI design that fits unobtrusively into user routines. One notable example of this is how care teams referred in our field study to "red" and "yellow" alarms, based on color coding of alarm states in their bedside monitoring system, suggesting an obvious color scheme for our interface.

Design Goal 5: Enable Trust Calibration through Transparency & Uncertainty Communication. Building trust in the AI-CDSS is essential for its effective use in clinical settings [27, 53]. It is important to ensure that users understand the system's limitations, acknowledging that the model and algorithm are not infallible [11]. A lack of appropriate trust calibration can hinder adoption and ultimately undermine the effectiveness of the proposed solution [27]. Communicating model uncertainties, arising from inherent imperfections and missing data (which is common in this setting) [26, 63], along with providing transparency into the reasons behind AI risk predictions [34, 53, 59], is essential to ensure users can trust the system effectively and make informed decisions about when to rely on AI outcomes and when to exercise caution.

5 Design of The HEART Interface

To address our design goals and explore methods to visualize risk score uncertainty, we developed a two-view interface for monitoring patient risk levels. The *Aggregate View* serves as an overview of all unit patients, while an *Individual Patient View* provides comprehensive information for individual patient assessment. An overview of the interface is shown in Figure 1. Interface design proceeded in parallel with development of the backend system, including predictive algorithms. To achieve later integration of the front and back-end components, both components were designed to meet the following constraints: (1) The AI is used to predict adverse events after surgery, starting when the patient is admitted to the CT-ICU. (2) The data updates infrequently (e.g., on a 15 min schedule). (3) Back-end databases and hospital patient tracking systems automatically add and remove patients to the *HEART* interface as patients enter or leave the CT-ICU. Since completing this design study, the interface has been connected to a working backend system that meets the above constraints and is being readied for deployment in the CT-ICU. The interface operates as a standalone web application that will be linked from within the electronic health record and made available on all bedside computers, wall displays, and at the nursing station. We anticipate that at the bedside, care teams will mainly use the Individual Patient View, focusing on information about the local patient. At the nursing station, where charge nurses maintain oversight over the whole unit, we anticipate that the care team will primarily use the Aggregate View, drilling down to the Individual Patient View when needed.

As part of the design process, we identified and addressed various usability issues throughout development. For example, early versions had confusing pinning, sorting, and filtering behaviors that we simplified based on feedback. Similarly, the presentation of short versus long length-of-stay outcomes proved confusing when displayed alongside other outcomes, prompting us to redesign how this information was organized and presented. Some usability discussions generated multiple design alternatives. While we do not exhaustively document all usability issues and fixes, we describe key design alternatives in this paper and explain which options were preferred by clinicians.

This section describes the final version of the designs (Figure 2-) *after* the expert walkthrough and a brief follow-up were completed. In Section 6, we describe some of the earlier variations before the final design, along with the expert feedback on those alternatives In this section, we occasionally include quotes from care team members collected during our iterative design process. Note that these quotes are distinct from the feedback gathered during the expert walkthrough.

5.1 Predictive Algorithms

We emphasize that the main contribution of this paper is the design of *HEART*'s user interface. The interface integrates outputs from real predictive models based on prior work by our co-author Jin and colleagues [32]. Thus, for context, we briefly describe these algorithms below; readers interested in their technical details should consult the original work.

HEART employs predictive models developed with XgBoost, a tree-based ensemble machine learning algorithm [14]. These models use a wide spectrum of electronic healthcare record variables for prediction, including demographics, patient risk factors, laboratory results, vital signs, clinical observations, and medications. Variables were categorized into two groups: static and dynamic variables. Static variables remain the same between hospital admission and CT-ICU admission, whereas dynamic variables represent features that change over time prior to CT-ICU admission. Detailed lists of static and dynamic variables are provided in the supplemental material. For dynamic variables, statistical metrics including number of measurements, mean, standard deviation, maximum, minimum, first and last measurement are calculated for the model input. The machine learning model predicts eight postoperative adverse outcomes in CT-ICU patients, including renal failure, stroke, prolonged

Table 1: Predictive algorithm performance. AUROC = area under the reciever operating characteristic curve.

Predicted Outcome	AUROC
Mortality	0.85
Stroke	0.73
Renal Failure	0.89
Prolonged Ventilation	0.85
Reoperation	0.73
Major Morbidity	0.81
Prolonged Length of Stay (PLOS)	0.82
Short Length of Stay (SLOS)	0.78

ventilation (defined as postoperative mechanical ventilation >24 hours), operative mortality, and length of stay, categorized as prolonged (>14 days) or short (<6 days). Model performance is reported in Table 1.

To compute risk score uncertainty based on missing values, we employed a Bayesian-based computational method [33]. Through this approach, Multivariate Imputation by Chained Equations (MICE) was applied to generate a distribution of missing values conditional on the available patient variables. The distribution of missing values represents their uncertainty. The distributions were then propagated through the ML models, generating a distribution of risk scores. Thus, when patient information is incomplete or missing, the model generates a range of possible risk levels in addition to the precise risk score. The primary assumption underlying the imputation strategy is that correlations exist among the features, allowing missing values to be inferred from observed variables under a Missing at Random (MAR) assumption. This assumption allows us to leverage observed variables to estimate missing values. For example, if a patient has elevated systolic blood pressure, their diastolic blood pressure is also likely to be elevated. MICE approximates the Bayesian posterior distribution of missing values by iteratively modeling each variable conditional on the others. This framework enables the use of different statistical models to estimate both the distribution and credible intervals of missing values given the observed data. However, MICE has several limitations. It typically relies on linear modeling approaches, such as generalized linear regression, to estimate conditional distributions. As a result, nonlinear relationships between variables may not be adequately captured. In addition, MICE can be computationally intensive, particularly when applied to large datasets or high-dimensional feature spaces.

5.2 Risk Categories and Color Coding

The predictive models estimate the probability of adverse events occurring, with the predicted probability defined as the risk score. Higher risk scores indicate greater likelihood of the patient developing adverse events. Aligning with care team requests in our field study [6], we defined risk categories based on risk score thresholds, assigning each category a color corresponding to risk severity. This approach aimed to simplify interpretation and provide care teams with an immediate understanding of the risk associated with each patient. From both our field and design studies, care teams habitually worked with categorized alert levels and expected that the predictive risk scores would be similarly categorized.

We engaged in extensive discussions with the CT-ICU care team and clinical researchers to determine how many colors they wanted to see (i.e., how many granular levels of risk for each event) and at which percentage thresholds the risk should shift to a higher or lower level. We also explored whether the risk thresholds should be consistent across all adverse events or if each event should have its own set of boundaries. This consideration arose from discussions with our health AI expert colleague, who pointed out that in some cases, certain risk predictions may show low numbers, but even minor changes could indicate a high risk. On the other hand, for other adverse events, higher numerical values often directly correlate with higher risk. Regarding both questions, no decisive answers emerged from the discussions themselves. Care teams either did not have a clear response, did not feel strongly about what would be best, or were unaware of any medical references that could serve as a standard approach. Therefore, we made some decisions around risk levels to use as a design probe, understanding that we would gain feedback to inform the number of risk levels and specific threshold values to be set at implementation time. For our design probe, we selected four risk categories: high risk and minimal risk, plus two moderate risk levels in between the two, to allow for more granular risk prediction. The color choice decision was influenced by clinicians' existing mental models of color coding in other monitoring systems, like the Philips monitor. The care team involved with the design study maintained "*[number and choice of the] colors align with clinical practices*" within the CT-ICU.

We selected *red* for high risk, *orange* for moderate-high risk, *yellow* for moderate risk, and *blue* for minimal risk. For the latter, we refrained from using green as is the norm in medical domain, to support green-red color blindness. These specific colors were chosen as a result of multiple design iterations, with the goal to balance visual urgency and readability while using softer tones to minimize alarm fatigue and eye strain. After discussions with clinical researchers on our team and the CT-ICU care team, we set a 40% risk threshold as the boundary for minimal (blue) risk, indicating that any value above this threshold warrants attention. To ensure consistency and avoid user confusion with varying thresholds for different adverse events, we adopted uniform risk ranges for values above this threshold while assuming that normalization techniques can be applied on the back-end.

We used these color categories as a basis to show risk scores and their uncertainty. Note that the uncertainty range could extend beyond the bounds of the predicted risk category, indicating that the risk category *might* shift with more data. These cases are critical, particularly when a patient categorized as minimal risk could actually be at moderate risk (akin to a false negative error). We referred to such instances as "borderline adverse events," where low risk might be moderate risk, or vice versa, due to incomplete data. The interface visualizes missing value uncertainty across different views, which we will describe separately for each view in the next two sections.

5.3 Aggregate View

The Aggregate view (see Figure 1 right) displays information in a structured table format, with each row representing a patient in the unit. Patients are sorted by overall risk level, with higher-risk

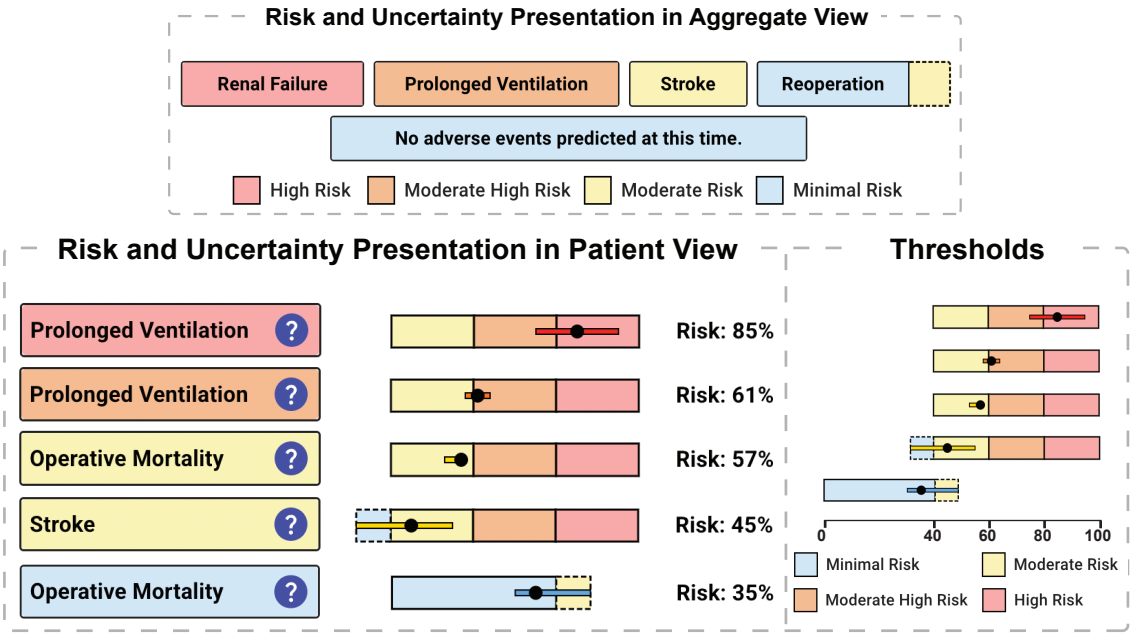
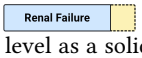
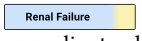


Figure 3: Top: Five distinct examples of adverse event presentations based on risk level in the Aggregate View. Note that minimal risk is not typically displayed, except in two cases: (1) when no high or moderate risks are predicted for any adverse events, a label is shown to indicate this (e.g., the label in the middle); and (2) when minimal risk falls within a borderline range, suggesting it might be moderate risk due to uncertainty (e.g., the *Reoperation* label in this image). Bottom Left: Five examples of adverse event visualizations from the Individual Patient View, including borderline cases where moderate-risk events may appear as minimal risk due to uncertainty. Bottom Right: Adverse risk category thresholds are displayed on a 0–100% scale for improved clarity.

patients appearing at the top, providing immediate attention to those with possible serious complications. In each row, the leftmost column includes a unique Patient IDentifier (PID), preceded by a colored circle that corresponds to the highest adverse event risk in the ensemble for each patient. The colored circle is a quick visual indicator that can allow clinicians to quickly identify high-risk patients at first glance. The 📌 icon can be used to pin any patient to the top of the list and non-pinned patients can be filtered. The care team emphasized simplicity in the aggregate view, noting that "*the aggregate view [should] be as simple as possible, [as a crowded interface risks exacerbating] alarm fatigue.*" They recommended to "*keep things simple and move detailed information to the patient-specific view.*" Accordingly, for each patient and within each row, a list of the predicted adverse events is shown as labeled and color-coded rectangular boxes, with color reflecting the risk severity (described in subsection 5.2). Minimal risk adverse events were excluded to reduce visual clutter, a need echoed by many care team members we talked to. For patients with only minimal risks, the system displays a blue label stating, "*No adverse events predicted at this time.*" Figure 3-Top shows examples from each possible representation.

To reduce visual clutter, we excluded uncertainty ranges from the Aggregate View. However, an important exception is made for borderline risks—cases where the uncertainty range of a minimal risk overlaps with moderate risk. These scenarios, which carry significant implications, are prominently highlighted and represent the only instances where minimal risk events are displayed in the Aggregate View. Because it was unclear what visual encoding mechanism would most effectively represent these uncertain predictions within the context of the color-coded rectangles, we implemented multiple visualization options, which we later assessed in our expert walkthrough. Here, we only show the designs included in the walkthrough:

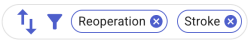
- **Dotted Boundary Visualization** 

. The interface shows the baseline predicted risk level as a solid box (such as minimal risk in blue), with a dotted outline extending to the right showing the potential range into higher risk categories. The size of the uncertainty box is fixed and is not dependent on uncertainty amount.
- **Gradient Visualization** 

. The interface displays the uncertainty range using gradient color that gradually transition blue into yellow. Similarly, the size of the uncertainty box is fixed.
- **Dotted Boundary with Range of Uncertainty** 

. In the third option, we represented the range of uncertainty using the length of a yellow box, with its size indicating how much the uncertainty range extends into moderate risk. Notably, the maximum uncertainty range displayed is capped at the moderate risk threshold. If the uncertainty range exceeds this threshold, we

recommend using a distinct design or color (e.g., grey) to signal that the model is too uncertain to make a reliable prediction [36].

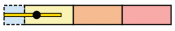
The interface offers interactive functions 

to help the care teams customize the display to monitor patients more efficiently. A filtering function allows care teams to select specific adverse events; e.g., a respiratory therapist might only be interested in looking at *prolonged ventilation*. For additional flexibility, users can choose to sort the filtered results either by the severity of the filtered events or by the patient's ensemble risk status.

5.4 Individual Patient View

The Individual Patient View (see Figure 1 left) provides detailed information and insights about an individual patient and can be accessed by clicking on a patient's name in the Aggregate View. This view includes three key components: adverse events with risk thresholds above 40% on the left, minimal-risk events with borderline uncertainty on the right, and a length-of-stay prediction visualization at the bottom. For consistency, adverse events within each group are sorted by risk score.

For each adverse event, the interface provides a visualization of both the prediction and its associated uncertainty. Each predicted risk score is represented by a colored box as in the Aggregate View. Adjacent to each adverse event name is a novel visualization that combines normalized risk and uncertainty range in a single widget. We designed a rectangular graphic divided by two thresholds into three ranges of equal size, with colors corresponding to one of three moderate-to-high risk categories  to illustrate risk boundaries at a glance.

Using the rectangular graphic as the background, we incorporated a black dot to indicate the exact normalized prediction point within the color-coded risk categories. This is paired with a line representing the range of uncertainty (*uncertainty band*), illustrating the potential outcomes given incomplete information  . The novelty lies in the combination of these three elements, enabling users to quickly assess the current risk's position within thresholds and understand how uncertainty might shift the risk across categories. As a design decision, the uncertainty band itself is represented in the same color as the category where the main prediction falls, only in a slightly darker shade for visibility. For moderate risk predictions where uncertainty can fall into a minimal risk level, the uncertainty band stretches leftward into an added dotted boundary region with constant size  . Figure 3-Bottom Left shows a few different examples of this widget based on different risk scores and uncertainty values.

On the right side of this widget, we displayed the actual value of the risk score. While we assumed these risk predictions were already normalized on the back-end, we considered the potential importance of showing both normalized and absolute (non-normalized) risk values. To explore this, we designed three alternatives for displaying the risk score:

- **Normalized Risk Score Label.** Only the normalized risk score as a percentage, in plain text (e.g., *Risk:* 83%), as in Figure 3.
- **Absolute Risk Score with Uncertainty Range.** The absolute risk percentage (actual non-normalized values), along with the uncertainty range (e.g., *Risk:* 2%[1.9 – 3%]), where 2% is the actual value considered high.

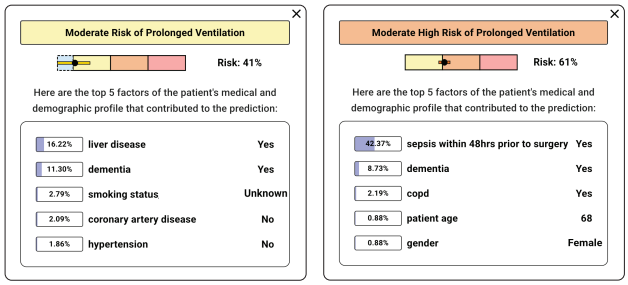
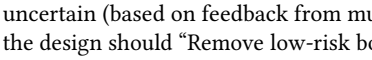
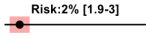


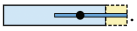
Figure 4: Two explanation panels for *prolonged ventilation*. Pressing the question mark next to an adverse event in the Individual Patient View 

opens this panel as a pop-up overlay. The top five (positive) SHAP values are displayed as the leading contributing factors to the risk assessment, along with their actual values. Note that, for demonstration purposes, these values are based solely on historical patient data; in practice, they can also incorporate live monitoring and updating information.

- **Absolute Risk with Visualization of Uncertainty Range** 

. A combination of option 2 with a line and black dot showing the exact position of the absolute risk, and a highlighted background color for the uncertainty range of the absolute value.

From the expert walkthrough, we received feedback from only one participant who preferred option 1. We retained these alternatives as optional views to accommodate different users, clinical contexts, and uncertainty preferences.

Minimal risk events are only shown when they are borderline uncertain (based on feedback from multiple clinicians, suggesting the design should "Remove low-risk boxes or indicators [i.e., those without uncertainty] from the interface to reduce visual clutter"). These are represented with a blue rectangle, a black dot indicating the exact risk prediction, a blue-colored uncertainty band, and an extended dotted yellow box to denote the borderline uncertainty  . Several alternative designs for the boundary box in the borderline region were considered (such as one similar to *Boundary with Range of Uncertainty* option in the Aggregate View), but they were ultimately excluded to avoid showing cases with uncertainty extending beyond moderate risk as those predictions are too unreliable.

For transparency, the Individual Patient View included the five most significant features contributing to each high-risk or borderline adverse event's risk score, analyzed by the Shapley Additive explanation (SHAP) value from the model [40] (a model-agnostic method to provide additive feature contribution with respect to the adverse event prediction for each patient). This information, initially hidden, could be accessed through a drill-down functionality by pressing the question mark button next to each adverse event label 

, triggering a pop-up display. We chose to display only positive SHAP values—those contributing to increased risk—since negative contributions were deemed irrelevant in this caregiving context. The explanations help care teams understand

model rationale behind predictions and potentially identify interventions to mitigate patient risks. Figure 4 shows example pop-up displays based on historic data. During the design study, the care teams strongly supported this transparency feature, with one clinician noting that "*the '?' is helpful to understand the model's decision-making process.*" Beyond acceptance, clinicians showed genuine interest in learning which patient attributes the model prioritized, and offered suggestions for additional features to include. While implementing these suggestions fell outside our design scope (requiring changes to data collection and model training), this enthusiasm demonstrates how clinicians value and actively engage with AI transparency features in healthcare interfaces.

Finally, we present a visualization for the patient's Length of Stay (LoS) prediction. According to STS [55] and other clinical guidelines, there are two major milestones for LoS in CT-ICU: 6 days and 14 days. Patients staying in the ICU for less than 6 days are considered low risk and are classified as having a short stay, while those staying beyond 14 days are categorized as high risk, indicating potential complications. Thus, LoS is an important indicator of a patient's recovery progress. It helps determine whether the patient is on track for full recovery and discharge or if their condition is more complex. To represent this information, we designed two visualizations, each presenting the exact prediction as an integer, along with the uncertainty and range of LoS based on important milestones in the CT-ICU:

- **Minimal Timeline Visualization.** In this design, day 6 and day 14 are marked on the timeline, with the patient's predicted length of stay indicated by a marker. The background color changes to blue, orange, or red, depending on which region the prediction falls into. The uncertainty range for the predicted days is shown in plain text (e.g., Range: [7–12] days) on hover.
- **Cohort Timeline Visualization.** This design features a similar set of categorical risk ranges to align with the four risk categories in the main risk score design. The boundaries are set at 3 days (for minimal risk), 3–6 days (for moderate risk), 6–14 days (for moderate-high risk), and 14+ days (for high risk). The actual predictive value for this patient is marked with a black dot, accompanied by an uncertainty band for the risk prediction. This design also includes a y-axis displaying a curve representing the distribution of stay lengths for a cohort of 500 past patients most similar to the current patient. This allows care teams to see how many historical patients stayed for each length of time. The combined view provides care teams with both historical patterns and precise predictions in a single visualization.

Figure 5 shows these two alternatives. The simpler design was shown to the experts who participated in the walkthrough study. The more complex visualization was developed and implemented after the walkthrough. In a follow-up discussion, one participant expressed strong approval of this design's potential.

6 Expert Walkthrough

Following the iterative design phase, we conducted a qualitative expert walkthrough study to gather clinical experts' feedback on the system design, including specific design elements, interactions, and design alternatives. The study provided an opportunity to gain input from a wider variety of clinical roles than was possible

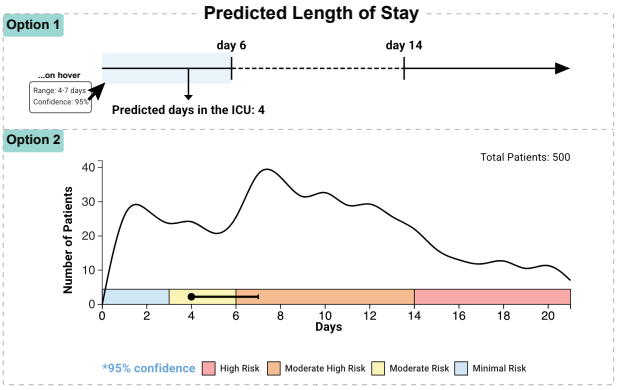


Figure 5: Two alternative visualizations for a patient's Length of Stay (LoS): Option 1 illustrates the Minimal Timeline Visualization, where the LoS prediction is categorized into three risk levels, shown with background colors based on two boundaries. The predicted LoS is marked by a downward arrow, with uncertainty displayed on hover. Option 2 presents the Cohort Timeline Visualization, using four risk categories consistent with other adverse events. The actual LoS prediction is shown with a black dot and uncertainty bands, while a distribution above the timeline represents LoS predictions for a cohort of 500 similar past patients.

throughout the design phase (where we primarily interacted with nurses and APPs), including experts from outside the collaborating CT-ICU.

The study was deemed exempt by the institutional review boards of the primary author's institution and the collaborating hospital. Verbal informed consent was obtained from participants following a verbal description of the study purpose. Since expert walkthroughs were recorded and transcribed, we obtained consent before recording began and offered participants the option to turn off their cameras. Two researchers attended each expert walkthrough session, with the senior researcher leading the discussions. Both researchers participated in the subsequent analysis and interpretation of results.

6.1 Expert Participants and Procedure

We recruited five healthcare professionals with diverse backgrounds in or related to the CT-ICU through snowball sampling [23]. This involved reaching out to our connections in the *MaineHealth Maine Medical Center*'s CT-ICU unit and asking them to connect us with other experts, internal or external to their hospital, who were interested and willing to participate. Two of the participants (**P3** and **P4**) were previously involved in the iterative user-centered feedback and design process, so their contributions to the expert study provided a holistic assessment of the interface's final design, functionality, and user experience. The other three participants, meanwhile, offered fresh perspectives on the interface's design and workflow.

Two participants (**P1**, **P3**) were current or former *Bedside Nurses*, two were physicians (**P2**, an *Intensivist*, and **P4**, a *Pediatrician*), and one (**P5**) was a *Physician Assistant*. All participants had direct

experience working in critical care units. In addition, **P3** had worked as a clinical researcher for a medical device manufacturing company, **P4** had expertise in computer science and had worked on medical algorithms, and **P5** had experience in surgery.

We conducted interviews via Zoom in November 2024, and had one or two participants per session, lasting between 30 to 60 minutes depending on number and schedules of the participants in each session. Two researchers were present during all sessions. At the beginning of each session, participants were given a brief introduction to the project and asked to describe their role. The researchers then demonstrated the the *HEART* interface, walking through its various functionalities, design choices, and alternatives, as well as the meaning behind its components and interactions. We described *HEART* as a reference tool that staff could consult throughout their shift, with the underlying predictions refreshing once or twice daily; most importantly before the morning shift begins. The system updates each morning to reflect new admissions and remove patients who were discharged or expired. The system was described to be available on monitors, workstations, and bedside stations throughout the CT-ICU.

Following the demonstration, we engaged the participants in an open-ended discussion, facilitated by the lead researcher, to gather feedback on the tool's components, areas for improvement, and what worked well. We explored three main areas: First, we asked participants to assess *HEART*'s usefulness for themselves and other CT-ICU team members, and to describe how it would support or hinder their daily workflows. Second, we gathered feedback on specific design decisions—from granular details like color choices and uncertainty representations to broader considerations like information density and interactions. Third, we asked participants to identify which parts of the interface they found most and least relevant for their specific healthcare roles, helping us understand role-specific needs. We also showed them design alternatives to identify which ones they preferred. While the researchers provided structure to the discussion, they prioritized participant-led dialogue, using follow-up questions to probe deeper into perspectives and clarify feedback. To counteract potential positive response bias, researchers explicitly encouraged participants to voice criticisms and concerns about the interface After the main interview session, we adjusted and updated some design elements and explored additional design variations based on the conversations. We invited the same participants for a follow-up feedback session in December 2024, and **P2** and **P3** accepted the invitation, providing valuable input that was used to update the designs.

With participants' consent, all interviews were audio-recorded, transcribed, and anonymized. Using In Vivo coding [22], one researcher extracted relevant excerpts from the interview transcripts and then applied affinity mapping to organize and identify emerging patterns and themes, which were then discussed extensively with another author.

6.2 Results

In the sections below, we summarize and highlight the key points of discussion, feedback, and opportunities for improvement. We begin with overall feedback on the interface and then drill specifically into visualization design, especially uncertainty visualization, the main contribution of our work.

6.2.1 General Feedback on The HEART Interface. Participants (**P1, P3, P5**) praised the interface for its simplicity and straightforward design. They emphasized the importance of making the interface accessible to users who are unfamiliar with AI, statistics, or advanced visualizations, important since technical literacy varies substantially in care teams [12]. Simplicity was seen as critical given their demanding schedules. In the information-saturated environment of the CT-ICU, where care teams are often overwhelmed by monitoring devices and extensive charting, *HEART* stood out as streamlined and intuitive, yet informative when compared to existing tools.

> "*[The Interface] is pretty straightforward. It's pretty simple. It's nice. [...] the more basic, the better.*" (**P5**)
>
> "*The EMR [Electronic Medical Record] is so visually overwhelming that this is so soothing to me.*" (**P1**)

P1 frequently used the term "user friendly" and "accessible"[3] to describe the interface. In their own words:

> "*I think this is a really great start [...]. It's accessible [...] that's the word that [keeps] coming to [my] mind. You have all of the most important stuff there [...] it's easily accessible. You can filter it down easily. You can get to what you need relatively quickly. I think anybody on the unit can understand this without really needing much explanation. [HEART] is definitely very user friendly [...]. In CT-ICU unit, this is something that will profoundly help us with our patient outcomes.*" **(P1)**

Participants appreciated efforts to keep the interface simple and uncluttered. For example, the decision not to display low-risk status for an adverse event unless the patient is borderline high-risk due to data uncertainty was well-received. The Aggregate View was seen as minimal yet highly informative, enabling users to gain valuable insights at a glance without feeling overwhelmed (**P1, P3**).

> "*If this was something that was on at the nurses station, and I'm coming in after a couple of days off, and I don't know what's happening on my unit. I can immediately tell [...] what the acuity level is on my unit [...] how many patients are kind of like what we would call hotspots.*" (**P1**)

However, it was noted that it would be worthwhile to optimize a version of the Aggregate View as a non-interactive overview for wall-mounted screens.

6.2.2 Risk Score and Uncertainty Visualization. The visualization of risk scores and their uncertainty in both Aggregate and Individual Patient views is a key contribution of our work and was a large focus of the walkthrough sessions. Participants generally supported using color encoding for categorized risk levels. They reported that they could instantly recognize the colors and assess the criticality

[3]The term *accessible* was primarily used by the participant to describe the availability of information and its ease of access, rather than referring to the broader design principles of accessibility for users with disabilities.

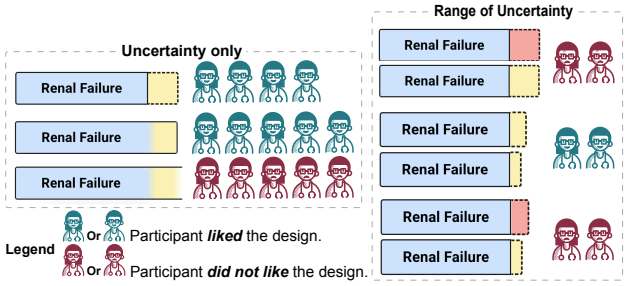


Figure 6: Participants were asked about their preferences for various design alternatives to represent borderline adverse events, particularly in the Aggregate View, where uncertainty bands are not displayed. Each icon in the figure corresponds to one participant (out of the total five). Interestingly, participant opinions were unanimously positive or negative for each design alternative. The range of uncertainty variations were added after the formal interviews, prompting the team to re-invite the original participants for quick feedback. Of these, only two participants were available and responded. The two participants supported incorporating a notion of uncertainty into the original representations by adjusting the size of the borderline box. However, adding colors revealed another issue: the interface should avoid representing cases where model uncertainty is so high that it resembles a guess by the AI, as it could undermine care teams' reliance on and trust in the model over time.

of a situation without delay (**P1**), as it reduced the need to interpret numbers or mathematical notations.

> "*One thing I like about what I'm seeing is the colors [...]. We humans tend to relate or to respond to colors in a specific way. I like that you are choosing colors because I think that will get patients' or [care team's] attention.*" (**P2**)

There were extensive discussions about the choice of colors used in the interface. While generally supportive of the color choices, **P2** expressed surprise at the use of *blue* for minimal risk instead of the more traditional *green*, as they associated green with low risk and an "okay" status. However, they appreciated the reasoning once informed about the prevalence of green-red colorblindness. Despite this consideration, we discovered after the walkthrough that the current color scheme does not fully accommodate all color vision deficiencies; specifically, the orange and red used in the interface can be difficult to differentiate for some users. Reducing the number of risk level categories would be another way to address this concern.

One key focus of the walkthrough was on risk score uncertainty visualization. As part of the interview process, participants were shown various design variations to gather feedback on their preferences and reasoning. An important take-away was that uncertainty information on the Aggregate view should be kept to a minimum. One key feature discussed was the display of borderline adverse events in the Aggregate View. Figure 6 displays these alternatives and summarizes participant feedback. Overall, four participants (**P1–P4**) expressed positive feedback on both the dotted line Renal Failure and gradient Renal Failure designs, while **P5** exclusively preferred the gradient design over all options. When asked to choose, three participants (**P1–P3**) preferred the dotted line design, while one (**P4**) favored the gradient. We originally designed a third option, featuring an open-ended gradient Renal Failure to show continuity, was universally disliked due to its resemblance to an incomplete or malfunctioning interface.

After the interviews and based on the feedback, the research team explored new ways to visualize both borderline uncertainty and the range of that uncertainty in the Aggregate View. This involved adjusting either the size and color of the borderline uncertainty box or both Renal Failure as a way to represent the amount of uncertainty, with smaller widths contributing to less uncertainty and vice versa, and warmer colors contributing to uncertainty extending across more risk boundaries. Two participants (**P2, P3**) provided feedback during a follow-up session. While both supported the use of box size to indicate uncertainty range, incorporating color raised significant concerns Renal Failure. They noted that highlighting cases where the model's predictions are highly uncertain—ranging from minimal to high risk—could undermine user trust. Such uncertainty might give the impression that the model frequently "guesses," potentially eroding confidence in its reliability over time. Based on this feedback from participants, we suggest not displaying predictions in cases of high borderline uncertainty. Instead, such scenarios could be communicated through messages like, "`The model is too uncertain about this adverse event to make a prediction at this time,`" or by using a distinct color code, such as gray. Accordingly, the range of acceptable uncertainty for borderline risks in the Aggregate View should be limited to only reflect the amount of uncertainty leading to the next immediate adverse score (i.e., minimal risk that can possibly become moderate risk or "yellow").

The posthoc explanations (Figure 4) provided through the top five contributing features for any risk assessment received positive feedback. Participants emphasized the importance of showcasing this explanatory information, citing that clinicians would not merely accept a resolution without knowing the rationale (**P1, P2, P3**), mirroring findings reported in prior work [34, 53, 59, 60]. For example, **P1** compared the risk score combined with feature importance to how care teams themselves evaluate potential adverse events: *"These are the kind of risk factors that we look at in terms of how a patient will tolerate [...] an intervention"*. Moreover, participants unanimously supported making this information available only on-demand, as it might overwhelm users in most situations and use cases.

> *"I think that's a great feature, and I think it's excellent that you can collapse it, and then reveal it like that. That's very helpful."* (**P1**)
>
> *"I do really like the ability to look and see what the top predictors are. That is a very often asked question, and clinicians love it. You don't need to look at [the explanations]. Click if you want to, and if you don't care, you don't care!"* (**P3**)

Participants also appreciated the focus on displaying only the top N contributing features (five in our case), rather than all possible

features. For instance, **P3** remarked that showing a smaller subset was particularly helpful in cases where "*there's clearly 1 or 2 big [contributing factors]*", and that they "*don't need to see the 40 other [features] that contributed like* 0.4% *[i.e., were minimally relevant]*". We note that some participants expressed confusion about the source or origin of the data and features (**P5**). This is a potential concern with SHAP-based explanations that may require additional training or more explanatory language in the interface.

6.2.3 Feedback on the Length of Stay Visualization. Another feature of our interface that sparked discussion was the length of stay prediction visualization. This feature was particularly valued by participants in administrative or managerial roles, such as charge nurses and intensivists, who oversee the care unit and are responsible for resource allocation, like bed management and staff scheduling. For these roles, the tool provides actionable insights for long-term planning and operational decision-making. Prior research also identified potential for AI in supporting these administrative functions [60], which were not the focus of our work. However, insights from our iterative design process with CT-ICU staff revealed that most physicians, surgeons, and physician assistants (PAs) found this feature less relevant to their daily tasks. Their focus remained primarily on risk and uncertainty visualizations—a sentiment echoed by our participant PA (**P5**). **P2**, who manages a team of intensivists while also caring for patients, offered a perspective highlighting this distinction:

> "*In the day-to-day ICU care, I'm not that interested in that [length of stay] data, but as an administrator or as a leader, I am; because it definitely helps me to kind of prognosticate and see what is gonna be my needs [in] terms of the staffing and resource allocation.*" (**P2**)

P1 explained that nurse schedules in their unit are currently planned a week in advance based on subjective judgment of patient progress. They found the predicted length of stay visualization helpful as it provides a data-driven estimate of when patients might be discharged, potentially improving efficiency in staffing decisions. Additionally, **P1** suggested that the visualization could benefit specialists such as respiratory therapists, who manage patients across multiple units with diverse ventilation and oxygenation needs:

> "*I can [imagine this visualization] would be very, very helpful for the respiratory therapists [...]. They are often managing patients on like multiple units, and just seeing [to], who has different ventilation and oxygenation needs.*" (**P1**)

P1 also made an interesting comment that alerted us about the importance of choices of coloring and how length of stay predictions are presented:

> "*Sometimes patients are medically stable, but they [...] have to stay in the ICU because they're [...] awaiting a permanent placement pacemaker; in a case like that, they would be staying in the ICU longer for a reason that is not necessarily medical. And then some patients are in the ICU longer, because they are very, very sick. And we're kind of spinning our wheels on getting them out.*" **(P1)**

This insight reveals that some CT-ICU patients, though medically stable, remain in the unit for reasons unrelated to their immediate health risks, such as maintaining their stability while awaiting surgery or follow-up treatment. Two implications arise from this observation: First, incorporating such cases into predictive models as training data (for refining the models over time) without context might lead to inaccurate predictions; for instance, a future patient who is stable and on track to leave the CT-ICU on time might be marked as high risk for certain adversities (false positive). Second, not all patients with extended stays (e.g., beyond 14 days) should automatically be categorized as high-risk based on their length of stay. These implications highlight the need for the interface to support user feedback or journaling capabilities to document such edge cases; ultimately, improving model training, enhancing human-in-the-loop workflows, and leading to better, more personalized care.

Finally, **P3** and **P4** emphasized the importance of not only presenting length-of-stay predictions but also providing the underlying reasoning behind these predictions. **P4** proposed a pathway tracking visualization as an alternative, which would allow care teams to monitor and track critical milestones and events in a patient's care journey instead of focusing solely on days. Such a visualization could inherently convey the rationale behind predictions by embedding context within the timeline of the patient's progression and minimize/contextualize the prediction uncertainties.

> *"if you know there's a pathway that a patient has to take in order to get out [...] tracking the patient's progress along that pathway might be a valuable thing. Were they extubated at midnight? Yes or no [...] Were they walking the next morning? [...] Have they peed, yet have they pooped yet? [etc.]"* **(P4)**

7 Lessons Learned

Our 16-month design process, which included extensive collaboration with user experience designers, AI researchers, and domain experts working closely with a CT-ICU care team, proved both rewarding and challenging. The journey was marked by numerous obstacles and contrasting opinions that we feel compelled to share our insights with the broader research community. In this section, we reflect on the lessons learned, highlighting key takeaways and potential future directions. We focus first on uncertainty visualization, the primary design focus of this work, and then touch on other observations that reaffirm prior work. There remain significant gaps in the current AI-supported tools for critical care workflows, and addressing these gaps could lead to transformative improvements in the caregiving paradigm.

7.1 The Importance of Visualizing Risk Score Uncertainty

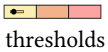
AI systems inevitably involve uncertainties and inaccuracies, which must be communicated to end users. However, in our review of the related literature during this project, we found that instances of visual AI-CDSS explicitly presenting uncertainty in risk score

predictions remain rare. The only closely-related examples we identified were *SepsisLab* [63] and Unremarkable AI [59]. One possible reason for this omission might be concerns among AI engineers that displaying uncertainty could lead users to dismiss the AI's recommendations entirely—for instance, perceiving the algorithm as unreliable or incapable. Nevertheless, Kompa et al. [36] argue that, much like physicians seeking second opinions from colleagues, an AI-CDSS should be capable of saying "I don't know" or withholding predictions when uncertainty is high. They strongly advocate that quantifying and effectively communicating uncertainties should be a top priority for AI systems in healthcare—a priority that is often overlooked.

We identified three primary types of uncertainties relevant to our context:

(1) **Prediction Uncertainty**: No AI algorithm is 100% accurate. Since AI models are trained on historical and available data, they can encounter cases unlike those in their training data. In such situations, the AI might misinterpret patterns or make incorrect assumptions, leading to prediction uncertainty.
(2) **Data Sparsity and Missingness**: Medical training data often contains sparse information and missing values. These gaps can arise from a variety of factors, such as incomplete patient histories or delayed or unperformed tests. In these instances, the AI's predictions are less reliable due to insufficient context, though they may improve as more data becomes available.
(3) **Statistical Nature of Predictions**: The predictions in our system are statistical and continuous (e.g., assessing the risk of an adverse event). When predictions fall in a "middle zone," they become inherently uncertain—akin to a 50-50 probability of a stroke, which offers little actionable insight.

While some of these uncertainties are universal to AI systems—e.g., (1)—others are domain-specific—e.g., (3)—and require careful study in their specific contexts. We urge the research and development community to identify and address the uncertainties inherent to their respective medical AI systems.

To address these uncertainties, our design included several strategies. For instance, we incorporated disclaimers and alerts communicating the general uncertainty of the model, such as: "⚠ `Use your knowledge and experience to monitor the patient. The algorithm predictions are not always accurate.`" Although this is not a comprehensive solution, it represents a meaningful first step. We also recognize and emphasize the importance of training and AI literacy education for ICU staff to help them understand the general uncertainties and limitations of these algorithms. Our primary focus was on uncertainties arising from missing data. As a solution to *minimizing* these uncertainties (as opposed to just communicating them), we extensively discussed human-in-the-loop approaches among the research team. One approach was to allow users to take control and input critical missing data as prompted by AI when filling missing values could improve the uncertainty in risk predictions (in our case, not all uncertainties warrant immediate attention, as resolving them might not always alter the course of action  , while other times a prediction could cross critical risk thresholds as more data becomes available, which could directly impact patient care ). Our results suggest that incorporating such features into AI systems may be valued by clinical teams.

7.2 Trust and Transparency Considerations for Designing AI-CDSS

In the CT-ICU at *MaineHealth Maine Medical Center*, while some care team staff advocate for integrating AI, they remain cautious. For users to effectively collaborate with AI in their decision-making process, especially in such a high-stakes context, they need a clear understanding of how the algorithm operates, as well as guidance on when to trust its predictions and when to rely on their own judgment [34, 53, 59, 60]. A particularly pressing issue is *algorithm aversion* among domain professionals, characterized by a reluctance to rely on algorithms perceived as unreliable [17]. Aversion can stem from various factors, including a lack of transparency, concerns over biased or problematic outcomes, and fears of overreliance on these algorithms for critical decision-making [1, 4]. A single misstep can have lasting consequences, especially since people are known to be less forgiving of AI's untrustworthy behaviors than other people's [16, 29]. Deployment and adoption studies by Henry et el. [27] and Sendak et al. [51] suggest that leadership endorsement, ongoing experience with an AI tool, involvement of care team members in the design process, and treating the AI model as a decision aid are essential ingredients to building trust in clinical settings.

Establishing trust requires transparency about both the "*why*" behind predictions (e.g., why a patient is at high risk for stroke) and the "*what*" (e.g., which factors in the patient's ongoing care influenced the decision). These explanations are particularly important during the early stages of interaction, when users may be skeptical and are building their relationship with the AI from a clean slate. Our design sought to build trust via several strategies: providing clear explanatory information, presenting uncertainty details about the risks, highlighting system limitations, and withholding predictions when the algorithm's confidence was too low. While explainability is only one of many factors in fostering trust and often requires meticulous refinement [44, 46], our design discussions with the care team consistently underscored its value, motivating us to expand explanation designs in future work. In our interface, we adopted feature-based explanations that is a commonly-used approach (e.g., [3, 53]), but other post-hoc explanation methods, such as counterfactual explanations [25], could be equally beneficial in this context.

Particularly, some participants expressed a need to understand what actions they could take to achieve different outcomes or reduce risk. This request ventures into the realm of diagnosis and prognosis, raising ethical concerns, as we aim to avoid direct care recommendations due to AI's susceptibility to errors. However, to address this need without crossing ethical boundaries, counterfactual post-hoc explanations [25] offer a promising middle-ground. These explanations illustrate correlations between outcomes and input factors. For instance, a counterfactual explanation might show how altering certain variables could decrease a risk prediction ($|f(x') - y'| < \epsilon$, here x represents input features and y denotes the outcome). When combined with interactive design elements like 'what if' scenarios [56], counterfactual explanations could enable users to explore

potential outcomes by manipulating variables; i.e., in our system, changes in which of a patient's health-record variables could alter any of the predicted adverse outcomes? For example, the care team could explore queries like "if the patient's heart rate would stabilize, would their high risk of prolonged ventilation decrease?"

7.3 Need for Standalone Tools to Predict and Track Patient's Length of Stay in the ICU

From our walkthrough study, we found that length-of-stay predictions were not considered very useful for patient care decisions, but would be very useful for managing workload and staffing. The ability to predict and track length of stay in the ICU is not just a technical challenge but a critical unmet need with far-reaching implications for care delivery and resource management. Through months of brainstorming and discussions with critical care clinicians and clinical software engineers during the design process, we envision a sophisticated tool dedicated to this task that could transform how ICU teams allocate beds, manage staff schedules, and plan interventions. Our findings reinforce prior observations that resource allocation is an important use case for AI in critical care settings [60]. Such a system could forecast patients' expected length of stay over days or weeks to support dynamic staffing. Feedback from our interviews suggests that to be most effective, a tool like this should offer explanatory insights that contextualize its predictions. For example, detailing whether a longer stay is due to a high-risk condition or the need for follow-up procedures or defining and tracking milestones in patient care to track patients' progress while in the unit. Moreover, AI explanations would also empower users to trust and act on predictions confidently.

7.4 Challenges of Designing for Clinical Care Units Due to Diverse Perspectives, Needs, & Workflows

Entering the design space, we anticipated diverse roles and expertise, but the CT-ICU setting (like other ICUs) is made up of personnel with varying responsibilities, years of experience, and team structures. Given the intense pace of patient care, an intuitive system is vital to minimize adoption time. In such environments, every second matters. The diversity within the team posed challenges, as contrasting feedback from various users required the design team to balance multiple concerns. Our approach included focusing on different user views, segmenting components into dashboard-like layouts, and simplifying technical details. Additionally, we encountered varied levels of technological comfort—some team members adapted quickly, while others resisted the new system.

Differences in roles also significantly influenced the type of feedback participants provided. Bedside nurses and administrative leaders offered more nuanced and detailed feedback on the interface design itself. In contrast, physicians and PAs were more focused on questions about the underlying models and algorithms; specifically, how they were trained, what data they used to base these predictions on, and other considerations that informed algorithmic risk and uncertainty predictions. These topics were somewhat outside the scope of the interface design and feedback sessions, but the research team strived to address these inquiries and relay any relevant information to the AI experts and model engineers. Given the interdisciplinary nature of AI-CDSS design, it is essential to identify the type of feedback needed and match it with the appropriate role. This ensures each group provides relevant insights based on their expertise.

Recruiting participants for feedback proved to be a significant challenge. Despite having partners at *MaineHealth Maine Medical Center* facilitating connections with the care team, participants' schedules were so packed that they had little time to sit down and reflect. Even on one occasion during the interview with **P5** (and a potential **P6**), they were responding to a *code blue* emergency, which not only caused them to be late to the meeting but also required **P6** to rush back to attend to the patient only two minutes after joining and ultimately get too busy to rejoin. **P5** made a humorous remark about the situation that might paint a better picture of the chaotic and unpredictable nature of the ICU:

> "*Sorry I was late to the meeting, and sorry I couldn't have more people here, but it's been a busy day here for us. So maybe your next project can be AI to predict how crazy our day will turn into!*"

To overcome these scheduling challenges, we made in-person visits to the CT-ICU during lunch hours and worked with charge nurses to connect with available staff. Another approach was to block calendar time for virtual feedback sessions, coordinated through team leaders who could bring along available colleagues with similar roles. Our snowball sampling approach was more breadth-focused and single-stage, closer to a Exponential non-discriminative snowball sampling [52]. We identified and leveraged participants in leadership roles who served as key connectors to the broader participant pool, but recruitment did not extend to a second wave where those participants would recruit others. One charge nurse proved particularly influential in our recruitment process. She possessed extensive knowledge of CT-ICU staff schedules, time-offs, and breaks, enabling her to identify optimal in-person visit times for participant engagement during our iterative design phase and to connect us with clinicians who had greater availability for expert walkthroughs. Relying on supportive leadership figures who understood the project was essential for participant access. While our snowball sampling remained single-stage (rather than cascading through multiple waves), it achieved considerable breadth by strategically engaging those in leadership positions. We learned early-on that conducting user-centered research in clinical settings requires flexibility from the research team, adapting to the needs and schedules of domain experts. We hope that the strategies we implemented will assist future researchers in effectively engaging critical care professionals.

8 Limitations

As a design study with a particular hospital's CT-ICU, our work is naturally limited in its generalizability. Nonetheless, the problem of representing risk scores with uncertainty information arises in many health applications, so our design contributions may have broader value. In addition, insights from our extended collaboration with critical care professionals may be useful to researchers wishing to engage in similar collaborations. We also note limitations of our expert walkthrough, including a small sample size and a single

coder for the qualitative analysis. Future steps in this research will include more comprehensive rounds of usability testing prior to deployment.

9 Conclusion

This paper outlined the interface design process for the *HEART* interface, a predictive AI system for the CT-ICU that was co-developed with clinical and AI experts over a 16-month iterative design process. *HEART* addressed the design challenge of communicating an ensemble of risk scores with associated uncertainty information for monitoring multiple patients simultaneously. The solution introduces novel visualizations that were well received by the CT-ICU care team. Our work contributes insight into uncertainty visualization design for AI-derived risk scores in a critical care application. The designs are now being integrated into a fully functional working system that will support continuous live data monitoring, in preparation for a deployment study and clinical trial at the collaborating hospital. Insights from this work offer concrete guidance and recommendations for Human-AI interaction design in clinical applications, especially critical care medicine.

10 Usage of Generative AI

In accordance with ACM IUI's policy[4], GenAI tools (specifically Claude) were lightly used for editorial support to occasionally improve the clarity and conciseness of author-written text. No GenAI was used in any research stage (system design, implementation, data analysis) or for generating new content. All ideas, analyses, and findings are the authors' original work.

Acknowledgments

This work was generously supported by Northeastern University's Impact Engines Award, "Healthcare Enabled by AI in Real Time (HEART)," and would not have been possible without support from an interdisciplinary set of experts and scientists. The authors would like to thank our collaborators from *MaineHealth Maine Medical Center* and *Nihon Kohden*; especially, we would like to thank Robert Kramer, Doug Sawyer, Jennifer Clement, Felistas Mazhude, Jennifer Low, Ben Tasker, Timothy Ruchti, Mohamed Elmahdy, Jessica Padykula, and Jules Bergmann. From Northeastern University, we thank Raimond Winslow for leading the broader HEART project, and Cliff Forlines, whose early involvement helped shape key interface design decisions. Most importantly, we would like to thank the members of the CT-ICU care team at *MaineHealth Maine Medical Center* and beyond, for taking time out of their busy schedules to participate in our design study and/or the expert walkthroughs. We are deeply grateful for their contributions.

References

[1] Mohamed Abouzahra, Dale Guenter, and Joseph Tan. 2024. Exploring physicians' continuous use of clinical decision support systems. *European Journal of Information Systems* 33, 2 (2024), 123–144.

[2] Mohamad Alameddine, Katie N Dainty, Raisa Deber, and William J Bill Sibbald. 2009. The intensive care unit work environment: current challenges and recommendations for the future. *Journal of critical care* 24, 2 (2009), 243–248.

[3] Amid Ayobi, Jacob Hughes, Christopher J Duckworth, Jakub J Dylag, Sam Gordon James, Paul Marshall, Matthew Guy, Anitha Kumaran, Adriane Chapman, Michael Boniface, et al. 2023. Computational notebooks as co-design tools: engaging young adults living with diabetes, family carers, and clinicians with machine learning models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[4] Sulemana Bankuoru Egala and Decui Liang. 2024. Algorithm aversion to mobile clinical decision support among clinicians: a choice-based conjoint analysis. *European Journal of Information Systems* 33, 6 (2024), 1016–1032.

[5] Amie J Barda, Christopher M Horvat, and Harry Hochheiser. 2020. A qualitative research framework for the design of user-centered displays of explanations for machine learning model predictions in healthcare. *BMC medical informatics and decision making* 20 (2020), 1–16.

[6] Carey Barry, Sydney K. Purdue, Mahsan Nourani, Clifton Forlines, Matthew S. Goodwin, and Melanie Tory. 2026. Needs and Barriers of Healthcare Teams to Implement Real-Time AI & Predictive Analytics in Cardiothoracic ICU: A Contextual Inquiry. (2026). Preprint available on OSF: https://osf.io/9esvk/overview?view_only=eb654f6af184470e8cd135c65b05d34c.

[7] Aditya Bhattacharya, Jeroen Ooge, Gregor Stiglic, and Katrien Verbert. 2023. Directive Explanations for Monitoring the Risk of Diabetes Onset: Introducing Directive Data-Centric Explanations and Combinations to Support What-If Explorations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 204–219.

[8] Serdar Bozyel, Evrim Şimşek, Duygu Koçyiğit, Arda Güler, Yetkin Korkmaz, Mehmet Şeker, Mehmet Ertürk, and Nurgül Keser. 2024. Artificial intelligence-based clinical decision support systems in cardiovascular diseases. *Anatolian Journal of Cardiology* 28, 2 (2024), 74.

[9] A Browne. 2023. Demographic characteristics and work experiences of physician scientists in the US: analysis of the 2022 National Sample Survey of Physicians (NSSP). Association of American Medical Colleges (AAMC). *DC: Association of American Medical Colleges* (2023).

[10] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 international conference on healthcare informatics*. IEEE, 160–169.

[11] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[12] Sarah Chang, Lucy Gray, Noy Alon, and John Torous. 2023. Patient and clinician experiences with sharing data visualizations integrated into mental health treatment. *Social Sciences* 12, 12 (2023), 648.

[13] Robert Chen, Vikas Kumar, Natalie Fitch, Jitesh Jagadish, Lifan Zhang, William Dunn, and Duen Horng Chau. 2015. explICU: A web-based visualization and predictive modeling toolkit for mortality in intensive care patients. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 6830–6833.

[14] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[15] Sharon Chiang, Robert Moss, Angela P Black, Michele Jackson, Chuck Moss, Jonathan Bidwell, Christian Meisel, and Tobias Loddenkemper. 2021. Evaluation and recommendations for effective data visualization for seizure forecasting algorithms. *JAMIA open* 4, 1 (2021), ooab009.

[16] Ewart J de Visser, Frank Krueger, Patrick McKnight, Steven Scheid, Melissa Smith, Stephanie Chalk, and Raja Parasuraman. 2012. The world is not enough: Trust in cognitive agents. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 56. Sage Publications Sage CA: Los Angeles, CA, 263–267.

[17] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *Journal of experimental psychology: General* 144, 1 (2015), 114.

[18] Yuhan Du, Anna Markella Antoniadi, Catherine McNestry, Fionnuala M McAuliffe, and Catherine Mooney. 2022. The role of xai in advice-taking from a clinical decision support system: A comparative user study of feature contribution-based and example-based explanations. *Applied Sciences* 12, 20 (2022), 10323.

[19] Daniel T Engelman, Walid Ben Ali, Judson B Williams, Louis P Perrault, V Seenu Reddy, Rakesh C Arora, Eric E Roselli, Ali Khoynezhad, Marc Gerdisch, Jerrold H Levy, et al. 2019. Guidelines for perioperative care in cardiac surgery: enhanced recovery after surgery society recommendations. *JAMA surgery* 154, 8 (2019), 755–766.

[20] Anthony Faiola, Preethi Srinivas, and Jon Duke. 2015. Supporting clinical cognition: a human-centered approach to a novel ICU information visualization dashboard. In *AMIA Annual Symposium Proceedings*, Vol. 2015. American Medical Informatics Association, 560.

[21] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The lancet digital health* 3, 11 (2021), e745–e750.

[22] Barney Glaser and Anselm Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

[4] https://iui.acm.org/2026/call-for-papers/

[23] Leo A Goodman. 1961. Snowball sampling. *The annals of mathematical statistics* (1961), 148–170.

[24] Navita Goyal, Connor Baumler, Tin Nguyen, and Hal Daumé III. 2024. The Impact of Explanations on Fairness in Human-AI Decision-Making: Protected vs Proxy Features. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 155–180.

[25] Riccardo Guidotti. 2024. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery* 38, 5 (2024), 2770–2824.

[26] Caitlin F. Harrigan, Gabriela Morgenshtern, Anna Goldenberg, and Fanny Chevalier. 2021. Considerations for Visualizing Uncertainty in Clinical Machine Learning Models. In *In Proceedings of CHI '21 Workshop: Realizing AI in Healthcare: Challenges Appearing in the Wild*.

[27] Katharine E Henry, Roy Adams, Cassandra Parent, Hossein Soleimani, Anirudh Sridharan, Lauren Johnson, David N Hager, Sara E Cosgrove, Andrew Markowski, Eili Y Klein, et al. 2022. Factors driving provider adoption of the TREWS machine learning-based early warning system and its effects on sepsis treatment timing. *Nature medicine* 28, 7 (2022), 1447–1454.

[28] Katharine E Henry, Rachel Kornfield, Anirudh Sridharan, Robert C Linton, Catherine Groh, Tony Wang, Albert Wu, Bilge Mutlu, and Suchi Saria. 2022. Human–machine teaming is key to AI adoption: clinicians' experiences with a deployed machine learning system. *NPJ digital medicine* 5, 1 (2022), 97.

[29] Robert R Hoffman, Matthew Johnson, Jeffrey M Bradshaw, and Al Underbrink. 2013. Trust in automation. *IEEE Intelligent Systems* 28, 1 (2013), 84–88.

[30] Matti Hokkanen, Heini Huhtala, Jari Laurikka, and Otso Järvinen. 2021. The effect of postoperative complications on health-related quality of life and survival 12 years after coronary artery bypass grafting–a prospective cohort study. *Journal of Cardiothoracic Surgery* 16, 1 (2021), 173.

[31] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C Ahn, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. Designing AI for trust and collaboration in time-constrained medical decisions: a sociotechnical lens. In *Proceedings of the 2021 chi conference on human factors in computing systems*. 1–14.

[32] Qingchu Jin, Saeed Amal, Jaime B Rabb, Felistas Mazhude, Venkatesh Shivandi, Robert S Kramer, Douglas B Sawyer, and Raimond L Winslow. 2025. Development and Validation of Machine Learning Models for Adverse Events after Cardiac Surgery. *medRxiv* (2025).

[33] Qingchu Jin and Raimond L. Winslow. filed 08/2024. A Computational Framework for Patient-Specific Uncertainty Quantification on Healthcare AI models. U.S. Provisional Patent. Ser. 63/678.

[34] Annika Kaltenhauser, Verena Rheinstädter, Andreas Butz, and Dieter P Wallach. 2020. " You Have to Piece the Puzzle Together" Implications for Designing Decision Support in Intensive Care. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1509–1522.

[35] Sara Klüber, Franzisca Maas, David Schraudt, Gina Hermann, Oliver Happel, and Tobias Grundgeiger. 2020. Experience matters: design and evaluation of an anesthesia support tool guided by user experience theory. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1523–1535.

[36] Benjamin Kompa, Jasper Snoek, and Andrew L Beam. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine* 4, 1 (2021), 4.

[37] Katarzyna Kotfis, Aleksandra Szylińska, Mariusz Listewnik, Marta Strzelbicka, Mirosław Brykczyński, Iwona Rotter, and Maciej Żukowski. 2018. Early delirium after cardiac surgery: an analysis of incidence and risk factors in elderly (> 65 years) and very elderly (>) 80 years) patients. *Clinical interventions in aging* (2018), 1061–1070.

[38] Jeremias Kuge, Tobias Grundgeiger, Paul Schlosser, Penelope Sanderson, and Oliver Happel. 2021. Design and evaluation of a head-worn display application for multi-patient monitoring. In *Proceedings of the 2021 ACM Designing Interactive Systems Conference*. 879–890.

[39] Dina Levy-Lambert, Jen J Gong, Tristan Naumann, Tom J Pollard, and John V Guttag. 2018. Visualizing patient timelines in the intensive care unit. *arXiv preprint arXiv:1806.00397* (2018).

[40] Scott Lundberg. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* (2017).

[41] Felistas Mazhude, Robert S Kramer, Anne Hicks, Qingchu Jin, Melanie Tory, Jaime B Rabb, Mahsan Nourani, Douglas B Sawyer, and Raimond L Winslow. 2024. Predictive Analytics in Cardiothoracic Care: Enhancing Outcomes with the Healthcare Enabled by Artificial Intelligence in Real Time (HEART) Project. *Journal of Maine Medical Center* 6, 2 (2024), 11.

[42] Tim Miller. 2023. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support using evaluative ai. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency*. 333–342.

[43] Sina Mohseni, Niloofar Zarei, and Eric D Ragan. 2021. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 11, 3-4 (2021), 1–45.

[44] Mahsan Nourani, Samia Kabir, Sina Mohseni, and Eric D Ragan. 2019. The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 7. 97–105.

[45] Mahsan Nourani, Joanie King, and Eric Ragan. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, Vol. 8. 112–121.

[46] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 340–350.

[47] Richard Osuala and Ognjen Arandjelović. 2017. Visualization of patient specific disease risk prediction. In *2017 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE, 241–244.

[48] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[49] Muhammad Rafiq, Pamela Mazzocato, Christian Guttmann, Jonas Spaak, and Carl Savage. 2024. Predictive analytics support for complex chronic medical conditions: An experience-based co-design study of physician managers' needs and preferences. *International Journal of Medical Informatics* 187 (2024), 105447.

[50] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. 2012. Design study methodology: Reflections from the trenches and the stacks. *IEEE transactions on visualization and computer graphics* 18, 12 (2012), 2431–2440.

[51] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. " The human body is a black box" supporting clinical decision-making with deep learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 99–109.

[52] Julia Simkus. 2023. Snowball sampling method: Techniques & examples. *Simply Psychology.[Veebileht] Https://Www. Simplypsychology. Org/Snowball-Sampling. Html (20.05. 2024)* (2023).

[53] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M Kahn, and Adam Perer. 2023. Ignore, trust, or negotiate: understanding clinician acceptance of AI-based treatment recommendations in health care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–18.

[54] Ibrahim Sultan, Valentino Bianco, Arman Kilic, Tudor Jovin, Ashutosh Jadhav, Brian Jankowitz, Edgar Aranda-Michel, Michael P D'angelo, Forozan Navid, Yisi Wang, et al. 2020. Predictors and outcomes of ischemic stroke after cardiac surgery. *The Annals of thoracic surgery* 110, 2 (2020), 448–456.

[55] Christina M Vassileva, Sary Aranki, J Matthew Brennan, Tsyoshi Kaneko, Max He, James S Gammie, Rakesh M Suri, Vinod H Thourani, Stephen Hazelrigg, and Patrick McCarthy. 2015. Evaluation of the Society of Thoracic Surgeons online risk calculator for assessment of risk in patients presenting for aortic valve replacement after prior coronary artery bypass graft: an analysis using the STS adult cardiac surgery database. *The Annals of thoracic surgery* 100, 6 (2015), 2109–2116.

[56] James Wexler, Mahima Pushkarna, Tolga Bolukbasi, Martin Wattenberg, Fernanda Viégas, and Jimbo Wilson. 2019. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 56–65.

[57] Siyi Wu, Weidan Cao, Shihan Fu, Bingsheng Yao, Ziqi Yang, Changchang Yin, Varun Mishra, Daniel Addison, Ping Zhang, and Dakuo Wang. 2024. CardioAI: A Multimodal AI-based System to Support Symptom Monitoring and Risk Detection of Cancer Treatment-Induced Cardiotoxicity. *arXiv preprint arXiv:2410.04592* (2024).

[58] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. Harnessing biomedical literature to calibrate clinicians' trust in AI decision support systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[59] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. 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*. 1–11.

[60] Nur Yildirim, Susanna Zlotnikov, Deniz Sayar, Jeremy M Kahn, Leigh A Bukowski, Sher Shah Amin, Kathryn A Riman, Billie S Davis, John S Minturn, Andrew J King, et al. 2024. Sketching ai concepts with capabilities and examples: ai innovation in the intensive care unit. In *Proceedings of the 2024 CHI conference on human factors in computing systems*. 1–18.

[61] Nur Yildirim, Susanna Zlotnikov, Aradhana Venkat, Gursimran Chawla, Jennifer Kim, Leigh A Bukowski, Jeremy M Kahn, James McCann, and John Zimmerman. 2024. Investigating why clinicians deviate from standards of care: liberating patients from mechanical ventilation in the icu. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–15.

[62] Hubert Dariusz Zając, Jorge Miguel Neves Ribeiro, Silvia Ingala, Simona Gentile, Ruth Wanjohi, Samuel Nguku Gitau, Jonathan Frederik Carlsen, Michael Bachmann Nielsen, and Tariq Osman Andersen. 2024. " It depends": Configuring AI to Improve Clinical Usefulness Across Contexts. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 874–889.

[63] Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M Padilla, Jeffrey Caterino, Ping Zhang, et al. 2024. Rethinking human-ai collaboration in complex medical decision making: A case study in sepsis diagnosis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–18.

[64] Jieqiong Zhao, Yixuan Wang, Michelle V Mancenido, Erin K Chiou, and Ross Maciejewski. 2023. Evaluating the impact of uncertainty visualization on model reliance. *IEEE Transactions on Visualization and Computer Graphics* (2023).