The Importance of Contextual Information in Clinical Decision Support: A Neonatal-Sepsis Case Study

Niels Helmantel[1,2], Hanna Hauptmann[1], Daniel Vijlbrief[2] and Richard Bartels[3,4,*]

[1]Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands
[2]Department of Neonatology, Wilhelmina Children's Hospital, UMC Utrecht, Utrecht, The Netherlands
[3]Julius Center for Health Sciences and Primary Care, UMC Utrecht, Utrecht, The Netherlands
[4]AI and Digital Innovation, UMC Utrecht, Utrecht, The Netherlands

Abstract

Machine-learning (ML) applications hold great promise in clinical practice, but user adoption remains challenging. Insights from human-computer interaction (HCI) show that many applications face critical user interface issues, notably the absence of contextual patient information alongside ML predictions. Our research attempts to bridge this gap by focusing on the design of an explainable interface for a sepsis prediction model in the neonatal intensive care unit of the Wilhelmina Children's Hospital. The interface presents contextual patient information in a centralized view, supporting healthcare providers in decision-making by facilitating the interpretation and validation of model predictions. First, user needs were identified and translated into requirements. Next, a solution was created through an iterative design processes. Finally, the effectiveness of the design was evaluated through a task-based think-aloud study where healthcare providers interacted with and provided feedback on the interface. The results demonstrate that for the sepsis algorithm to be an effective decision support, relevant contextual information is required by clinicians to interpret and validate ML predictions. The study highlights the importance of interaction design in creating valuable ML applications and contributes to designing trustful and purposeful interfaces for ML-based systems in clinical practice.

Keywords

human-computer interaction, machine learning, user interface, XAI, sepsis

1. Introduction

Diagnosing and treating medical conditions within the intensive care unit (ICU) poses an intricate challenge, complicating the delivery of appropriate care. Caregivers have to analyse and piece together large amounts of information amidst cognitively demanding and time-sensitive situations. Despite the substantial capacity and expertise of healthcare providers, critical information occasionally evades notice due to the constraints of human memory, cognitive biases and inefficient communication [1]. These limitations can results in major consequences, potentially subjecting patients to enduring harm or even fatality. Furthermore, the healthcare sector is grappling with a growing staff shortage, leading to a diminished workforce available to oversee patients. Consequently, there's a heightened likelihood of failing to detect deterioration in a patients' condition. The combination of machine learning (ML) and the vast amounts of data stored in the electronic health record (EHR) presents a potential solution for these challenges [2]. ML excels in discerning trends and patterns in large amounts of data such as vital signs, laboratory findings, and various clinical parameters that could potentially signal early stages of patient deterioration. By continuously monitoring these parameters, care providers can promptly receive notifications about potential risks, allowing timely start of treatments.

Despite the enormous enthusiasm surrounding the integration of ML into the healthcare domain, many applications are not adopted due to issues such as inadequate interpretability or usability of ML models [3, 4]. In healthcare decisions have a major impact on patients' well-being and when a healthcare provider is unable to grasp the process that led to a prediction, then it becomes difficult

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus
nielshelmantel@gmail.com (N. Helmantel); h.j.hauptmann@uu.nl (H. Hauptmann); D.C.Vijlbrief@umcutrecht.nl (D. Vijlbrief); r.t.bartels-6@umcutrecht.nl (R. Bartels)
0000-0002-6840-5341 (H. Hauptmann); 0000-0002-2682-7386 (D. Vijlbrief); 0000-0001-7214-8029 (R. Bartels)

to determine the predictions its reliability. This issue becomes even more prominent in instances of errors in the prediction or unforeseen outcomes. The majority of research aimed at enhancing the interpretability of ML models, known as explainable AI (XAI), focuses on technical challenges [5]. Nonetheless, the effectiveness of an explainable model is heavily determined by the end user, the intended goal, and the context of use [6]. This underscores the necessity of user-centric XAI research, especially as an increasing number of machine learning models are being implemented in clinical practice. Human-computer interaction (HCI) researchers have developed frameworks and guidelines to design explainable ML models. Nonetheless, two-thirds of XAI interfaces in healthcare have critical problems [4]. One contributing factor is the absence of contextual patient information alongside ML predictions [6, 7]. Healthcare providers underscore the importance of contextual information to comprehend and trust predictions. For instance, one study found that historical events are needed as evidence to determine whether a prediction is reliable [8]. In a similar vein, nurses who received sepsis alerts felt uncomfortable making decisions because sufficient patient information was lacking [9], and electrophysiologists in some cases relied on contacting patients for more information to interpret risk predictions [10]. Despite a large body of research highlighting the importance of contextual information in XAI interfaces, limited research has been done on what type of contextual information is required.

In this study we aim to scope the contextual information required in the use case of a neonatal late-onset sepsis (LOS) prediction. LOS is defined as sepsis presenting $\geq$ 72h after birth and is a major cause of morbidity and mortality. In the most vulnerable preterm patients, the prevalence can be as high as 50% with mortality rates upto 35%. In addition, LOS survivors have an elevated risks of severe neurodevelopmental impairment [11]. While timely treatment is of utmost importance, LOS presents itself with nonspecific symptoms such as instabilities in heart- and respiratory rate, oxygen saturation and temperature. To facilitate early diagnosis researchers at the neonatal intensive-care unit (NICU) of the Wilhelmina Children's Hospital (WKZ) developed an early-warning system that aims to predict LOS before initial clinician suspicion, allowing healthcare providers to start treatment earlier [12]. Similar algorithms have been introduced by other groups [13, 14]. The algorithm by van den Berg et al. [12] was run in shadow-mode, meaning that a realtime implementation of the model was technically realized, but care providers were not exposed to the predictions by the algorithm. The predictions of the model can have a major positive or negative impact, as they can affect a clinician's decision to start or delay treatment, i.e. administering antibiotics. Because of this, healthcare providers need sufficient information about both the model and the patient to interpret and validate predictions. This is especially important because the algorithm is *imperfect*, with a precision and recall of about 5% and 60%, respectively. For example, if the model predicts an increased risk of LOS based on an elevated heart rate and the infant has just been administered medication then that may be reason to reject the alarm. This is consistent with studies that found that healthcare providers need contextual patient information, such as vital parameters, laboratory results and patient history to interpret risk scores [6, 9, 15]. Moreover, as the condition of LOS patients can deteriorate rapidly, it is crucial that predictions can be easily interpreted and decisions can be made quickly by a variety of healthcare providers. However, patient information is scattered across various sources, requiring healthcare providers to search for information in many different places and to do a lot of clicking [16, 17]. On the contrary, research suggests that information synthesis, for instance through synchronized interactive timelines, aids interpretation [16, 18]. In the context of a prediction model for necrotizing entercolitis Gephart et al. [17] found that healthcare providers require a centralized location for contextual information to interpret predictions.

This paper investigates how contextual patient information should be incorporated into an XAI interface for the late-onset sepsis prediction model from Berg et al. [12] with the research questions:

RQ1 What specific contextual information do healthcare professionals require to interpret LOS predictions?

RQ2 How can contextual information be best presented in an XAI interface to facilitate the interpretation of LOS predictions?

RQ3 What is the impact of a dashboard with contextual information on trust, reliance and decision-making for healthcare professionals?

2. Materials and Methods

The nine-stage design methodology from Sedlmair et al. [19] was used to develop an XAI interface for the LOS model of van den Berg et al. [12]. It consists of three top-level phases: precondition, core and analysis, subdivided further into stages. Although the framework is presented as a sequence, it does not mean each previous stage must be entirely finished before the next begins. Many phases overlap, and the process involves significant iteration. As each step is carried out, new information might emerge, refining previous stages. A brief overview of the stages is provided below, for further details see Sedlmair et al. [19]. The core four stages are: discover, design, implement, and deploy.

2.1. Precondition

Before the core phase, the preconditions need to be concluded. The initial stages of learn, winnow, and cast revolve around getting the visualization researcher (N. Helmantel) ready for the task and identifying and selecting collaborative partnerships with domain experts. The learn stage provides knowledge of the visualization literature, including visual coding and interaction techniques, design guidelines and evaluation methods. This knowledge provides an important foundation for later stages. For instance, it helps with data and task abstraction in the discovery phase and offers the researcher the ability to distinguish between good and bad ideas in the design phase. Next, the winnow phase focuses on selecting promising collaborations. This involves meeting with a large number of potential partners and gradually selecting a few based on careful consideration. Selection criteria can be practical, intellectual as well as interpersonal, e.g. the time available between two parties. Finally, during the cast stage collaborator roles are identified. The two most important roles are the front-line analyst, i.e. the domain expert end user of the product, and the gatekeeper who can approve or reject the project.

2.2. Discover

In the discovery phase, the problem will be characterized by talking to and observing different domain experts so that the domain-specific insights can be abstracted. This process is iterative: the researcher asks questions to the domain expert, the researcher abstracts and asks for feedback on the abstraction. A good abstraction ensures that the results from this design study can be used in other domains, and provides an understandable description of the domain for a visualization audience. This stage is used to gather requirements on the design. These requirements were classified into four priority categories. *Must-haves* are critical requirements. *Should haves* are important, but not necessary for the prototype to be validated. *Could haves* are desirable, but not necessary. These can be included depending on available time and resources. *Won't haves* are least important and will not be included. An observational study and user interviews were performed to identify and prioritise requirements.

Observational Study Half a day was dedicated to observing care providers on the NICU ward. Insights were gathered pertaining to decision-making, collaboration, and information exchange. This included, following a neonatologist during their morning routine, which comprised patient visits and offering guidance to fellow care providers. Nurses were observed while delivering care, providing an opportunity to pose questions for clarification of their actions. And the neonatologist demonstrated the utilization of the EHR (MetaVision) for patient administration. Finally, the researcher participated in the multi-disciplinary patient rounds where decisions were made regarding the treatment of challenging cases. Throughout the observation, detailed notes were taken and later expanded upon for comprehensive documentation.

User Interviews Semi-structured interviews were conducted to identify user needs. Interviews were conducted independently and lasted approximately 30 minutes. They began with a brief introduction of the study purpose. It was made clear to the participant that the interview would be recorded, the information would be kept confidential and the interview could be stopped at any time without having

to provide a valid reason. Participants were then asked to sign a consent form (Appendix A). The interview delved into the intricacies of sepsis diagnosis, the methods employed for information analysis, and the perspectives of healthcare providers concerning the introduction of a sepsis prediction algorithm to their ward. The interview outline can be found in Appendix B. A second round of interviews was conducted to identify the exchange of information between caregivers. The outline of the interview can be found in the Appendix C.

2.3. Design

In the design phase, the data abstractions, visual encodings and interaction mechanism will be generated and validated. Building on the requirements identified in the discover phase and the XAI and visualisation literature, multiple solutions were generated. A common pitfall is to choose a solution too early. This can be avoided by generating a wide selection of solutions. Solutions can be incrementally refined using design principles and guidelines. The proposed solutions should be presented to domain experts for discussion, such as in the form of paper mock-ups, data sketches or low-fidelity prototypes. The goal of the design cycle is to satisfy rather than optimize: while there is usually no best solution, there are many adequate solutions.

Low-Fidelity Prototypes Low-fidelity prototypes are simple, rough and often quick designs of a product or interface. The focus is on the concept and not the detailed graphics. This ensures that concepts can be tested quickly to validate hypotheses and gain valuable insights. A low-fidelity prototype can be created in a variety of ways. This study used Figma[1], a web application for designing clickable interfaces. This allows users to use the interface in a real-world manner, resulting in practical insights. The low-fidelity prototypes were designed based on Munzner [20] and XAI literature [6, 21].

Domain experts were asked to examine the low-fidelity prototypes, aiming to confirm the alignment of the designs with their requirements. Participants participated in the interviews either in person in a private room or, if in-person attendance was not feasible, the session was held online via Microsoft Teams. For these sessions, the low-fidelity designs were made available on paper (in-person session) or via Figma (online). The answers to the interview questions were recorded on a mobile device. The interviews lasted a total of 30 minutes per participant. The interview started with a general introduction to the topic, explaining the purpose of the study and the expectations of the interview. Participants were informed on the formalities and were asked whether they agree to an audio recording before signing the informed consent form. Next, participants evaluated the interfaces on whether the information displayed was sufficient for interpreting predictions, and whether the presentation of the information was both understandable and user-friendly. They were asked which visualizations they found most understandable and useful (i.e., visual encoding and interaction mechanisms) for estimating the risk of sepsis. Finally, participants were asked to compose a dashboard to their liking. They could use printed versions of the components or draw a lay-out by using pen and paper. During the interview, participants were encouraged to use pen and paper to draw, for example, alternative visualisations. The session ended with the researcher summarising the main insights, which participants could then confirm. The full interview outline can be found in Appendix D.

In addition to the domain experts, usability experts were asked to share one positive and one negative comment for each of the component.

Finally, all information was analyzed. The audio recordings were analysed, findings noted and afterwards the audio recording was deleted. Duplicate findings were extracted and the remaining findings were sorted from most important to least important for the task at hand. Feedback not relevant to the dashboard was removed from the list and saved for future work. Based on the feedback, a selection was made of the visualizations that participants found most understandable and useful, which were used for the high-fidelity prototype.

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

2.4. Implement

Once a design solution had been identified in the preceding phase, a high-fidelity prototype was created. A high-fidelity prototype is an advanced, detailed and interactive model of a product or system that accurately mimics the look, feel and functionality of the final product. This type of prototype is often developed using technologies and tools similar to those used to develop the final product, providing a realistic user experience. Using data from the design stage, a high-fidelity web application prototype was developed using JavaScript, HTML, and CSS. The interactive charts were developed using D3.js[2]. Subsequently, this prototype was tested with the envisioned end users. The evaluation assessed three primary variables: usability and comprehension, trust and reliability, and decision-making.

The System Usability Scale (SUS), a standardized questionnaire used to gauge the usability of an interface, was used to measure user satisfaction, as detailed in Appendix E [22, 23]. A higher SUS score correlates with enhanced usage and greater user satisfaction.

The high-fidelity prototype was further evaluated by conducting a task-based think-aloud study, a widely recognized method in usability research [24]. The qualitative nature of this study aims to measure the impact of the designed dashboard on trust and reliance, usability and decision-making. Participants were asked to articulate their thoughts aloud while performing tasks, providing important insights into their decision-making processes. Four distinct tasks emulated real-world scenarios. For privacy reasons, no identifiable patient data was used in this study. Instead, anonymized data (e.g. vital parameter timeseries, lab-tests, etc.) were used to create the four scenarios. Two situations included a sepsis alarm, with one being a true- and the other a false positive. In the other two situations there was no sepsis alarm, with one being a false negatives. Whether the (non) alarm was correct or not was deliberately hidden from the participant. Moreover, to control for any learning effects the sequence of tasks were shuffled between users according to the Latin-square method. The task specifications were:

Task 1: true positive An alarm was raised at 13:00 for a patient at high risk of sepsis. Are you concerned and how high do you estimate the risk?

Rationale task: the goal of this task is to find out what information they use for decision making and what strategies they use to come to a conclusion.

Task 2: false positive An alarm was raised at 17:00 for a patient at high risk of sepsis. Are you concerned and how high do you estimate the risk?

Rationale task: incorrect predictions can result in unnecessary interventions and treatments. The objective of this task is to determine if healthcare providers can identify false positives and to understand how they come to this conclusion.

Task 3: true negative No alarm was generated for the following patient. Are you concerned and how do you estimate the risk?

Rationale task: the goal of this task is to find out what information they use for decision making and what strategies they use to come to a conclusion.

Task 4: false negative No alarm was generated for the following patient. Are you concerned and how do you estimate the risk?

Rationale task: false negative predictions can lead to missed diagnosis or delayed treatments. This situation is unlikely to occur because healthcare providers will consult the dashboard when an alarm has been generated. However, it is still useful to know if care providers are able to detect false negative predictions based on the information in the dashboard and, importanty, how they arrive at this conclusion.

[2] https://d3js.org

Semi-structured interview sessions were held in-person in a private room, so that the evaluation could talke place without interruption. Interviews were recorded using a cell phone and screen recordings were captured using pre-installed software. The session started with an introduction of the researcher, an explanation of the research goal and what is expected from the participant during the session. Then participants were asked to introduce themselves, and asked for their age, sex, domain-expertise, and years of experience. Users were then asked whether they agreed to an audio and screen recording during the session while performing the tasks. It was also made clear that they could stop at any time and need not give a valid reason for doing so. Participants signed an informed consent form. After the introduction, the researcher provided background information about the sepsis prediction model, including why the model was developed in the first place and what variables were used to make a prediction. Furthermore, the researcher explained all components of the high-fidelity prototype and participants could ask any questions during the explanation. Then, participants were asked to take place behind a computer with the high-fidelity XAI interface already open. Before they started with the first task, participants were asked to think-aloud while doing the tasks and mention: what information they were looking at, what they were doing, what they liked and did not like. Furthermore, it was explained that there were no right or wrong answers and that no score was kept in the background. It was made clear to participants that they could always ask questions while performing the tasks, for example if they did not understand something on the dashboard. Then, participants were asked to start with the first task. After they were satisfied with their answer, the researcher asked how they estimated the risk and what actions they would perform based on that conclusion. To find out to what degree participants trusted model predictions, they were asked whether they agreed with the prediction and how confident they were in their decision. Participants were then asked what components on the dashboard were most valuable for estimating the sepsis risk to assess the effect of contextual information on trust and reliance. Here, contextual information is broken down into components such as patient information, vital signs, and feature contributions. After completion of all four tasks, participants were asked to complete the SUS questionnaire to evaluate the user-friendliness of the dashboard, whether there is sufficient information on the dashboard, and if they would have visualised it in a different way. Participants were thanked for their time and the recordings were stopped. The full interview outline is in Appendix F.

Audio-recordings were transcribed using Amberscript and analysed with NVivo (Lumivero) to code text and extract themes. The impact of contextual information on trust and reliability, and decision-making was examined after analysis of the transcriptions and screen recordings.

2.5. Deployment and Analysis Phase

In the final core phase, we put the tool to real-world use and gather feedback. As the sepsis tool is not yet applied in clinical practice this step was not performed. After the core phase, the analysis phase concludes the project. The analysis phase consists out of a reflection stage that can be used to confirm, refine, extend, or even propose new design guidelines and communicate them to the community.

3. Results

3.1. Winnow & Cast

The following roles were involved in this study. Note that a single person can have multiple roles.

Researchers The researchers involved in this project are the authors of this paper. They consist of experts in human-computer interaction (NH and HH), a neonatologist (DV) from the NICU of the WKZ, and a data scientist (RB). DV and RB were involved in the development of the algorithm (Berg et al. 2023) [12]. NH is responsible for designing and conducting the interviews.

Domain Experts The domain experts involved in this project are the following healthcare providers: neonatologists, NICU nurses, NICU physician assistants and Neonatal Fellows.

Usability Experts Master students from the Human-Computer Interaction program at Utrecht University were tasked with providing interface feedback in the design phase of the study.

3.2. Discover

Seven participants, two neonatologists (P1, P2), a Neonatal Fellow (P3), infectiologist (P4), and three nurses (P5, P6, P7) participated in the initial interviews. One participant was interviewed twice. In addition, requirements were gathered from XAI literature (L) [6, 7, 15, 25]. The interviews with care providers and related work yielded four themes: explanation, patient information, model feedback and collaboration, and alarm notification. The resulting requirements are detailed in Appendix G.

Care providers indicated reluctance to make decisions based on a prediction alone. They would like to know what the prediction is based on (req. 1.1) so that they can cross-validate it with available data about the patient. Additionally, they wanted to know the trend of the prediction (req. 1.2, 1.3). A high-risk patient coming from a very low risk prediction may be more alarming than a patient maintaining a moderately high risk score. There was disagreement about showing reliability metrics. Some care providers expressed a desire to know the likelihood behind a score and wanted information on sensitivity and specificity, whereas others found these metrics too challenging to understand. Since the majority of care providers did not want to see this on the dashboard, it was not included.

Contextual information was deemed key to facilitate informed decision making by care providers. It is needed to interpret and compare model risk scores with their own insights. In addition, they emphasized that it can enhance their confidence in model predictions. While the EHR contains a lot of contextual information, not all of it is relevant for interpreting sepsis risk scores. Care providers expressed a desire to have all relevant information consolidated in one central location, enabling them to act quickly and efficiently. Providers mentioned that it takes a significant amount of time to find to all relevant contextual information as it is scattered across multiple sources and hidden in long reports, resulting in instances where crucial details are overlooked and it takes longer to attend the patient. Centralizing all information relevant to interpreting sepsis risk scores would enable providers to act efficiently without spending time searching for scattered information across multiple sources.

During the interviews, care providers explained the diagnostic process of sepsis and all the relevant parameters. Together with a neonatologist, this extensive list of parameters was condensed to the most relevant ones. Interventions like antibiotic administration and respiratory support was assumed to be of importance for interpreting the risk scores. For example, respiratory support might affect oxygen saturation, a predictor used by the model, and thus could influence the risk score (req. 2.1). Additionally, it provides insights into the child's stability; if oxygen saturation is highly unstable without respiratory support, it could be a cause for significant concern. Furthermore, clinical symptoms play a role in evaluating the patient's condition and assessing the severity of the situation. The child's color and mobility are particularly important: if the child appears still, fatigued, and/or has a pale complexion, it is highly concerning (req. 2.2). Furthermore, laboratory results like white blood cell count and CRP values serve as indicators for infections in the body (req. 2.3). Vital functions provide indicators of the vital organs performance, with heart rate and oxygen saturation being the most critical ones (req. 2.4). These are also the input features for the prediction model. Healthcare providers expressed a desire to know the patient's weight and age (req. 2.6). Knowing that the child is extremely premature, treatment may be initiated earlier and a moderate risk scores could be more alarming.

The literature shows that providing feedback on a prediction has a positive effect on usability (req. 3.1). In addition, model developers can use this information to further refine the model. Healthcare providers wanted to know what actions were performed when an alarm was validated (req. 3.2).

Finally, care providers wanted alarms to be forced on them so non would be missed (req. 4.1).

3.3. Design Phase

A range of low-fidelity prototypes were created based on the identified requirements. The following presentation components were identified to meet the requirements: vital sign charts, feature importance chart, risk analysis, risk score, table with patient information, NICU unit overview section, and a feedback form. First, for each of the requirements, the data types were determined, including the attribute type (e.g., categorical, quantitative), ordering direction (e.g., sequential, diverging) and range.

Subsequently, based on XAI and visualisation literature [20], a set of relevant design solutions was created. Examples are provided in Appendix H.

The low-fidelity prototypes were evaluated by a total of nine subjects: six domain experts, one data scientist, and two usability experts. The domain experts consisted out of one Neonatal Fellow, one neonatologist, two nurses and two physician assistants. All care providers possessed expertise in diagnosing and treating neonatal sepsis. They exhibited diversity in terms of age and practical experience. The usability experts were second-year HCI M.Sc. students, providing both practical and academic insights. The main findings from the interview are discussed below.

Dashboard Composition Domain and usability experts agreed on the layout of the dashboard (Figure 1). They wanted all patients from a unit on the left side of the screen and more information about that specific patient on the right side of the screen. Domain experts thought the feature importance chart should not have a permanent location on the dashboard as the information can be easily deferred from the vital signs charts. Care providers thought the vital signs chart was more important than the risk analysis chart and should therefore be placed higher.

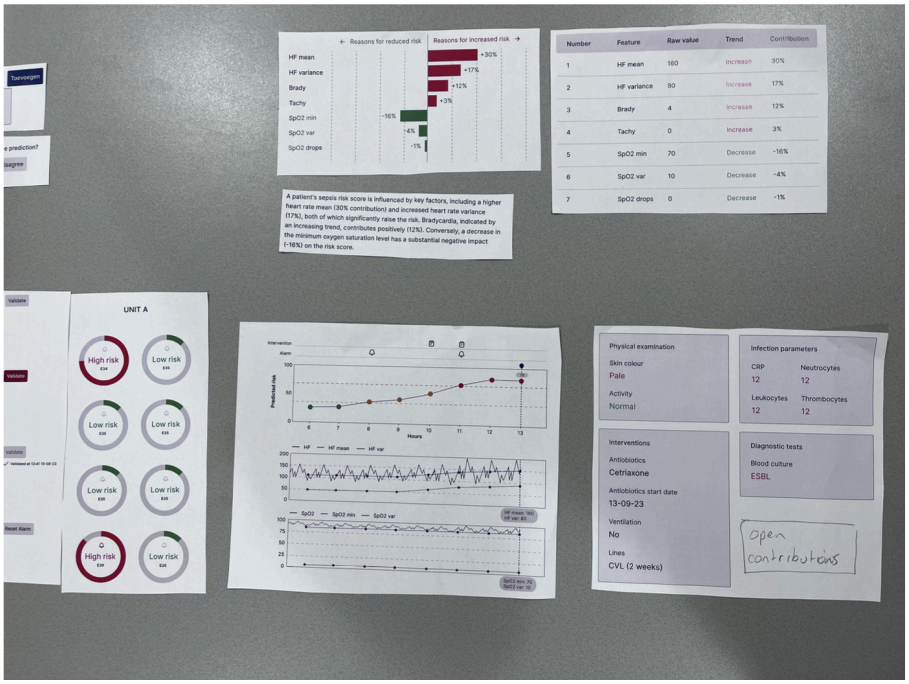


Figure 1: Dashboard layout composed by the participants. Participants preferred a patient overview on the left. Predictions and vital signs are shown at the bottom (middle) and model explanations at the top. General patient characteristics and history are placed in bottom right.

Vital Signs Care providers showed a preference for a line chart featuring multiple attributes in a single plot. They wanted to compare attributes within one consolidated view. They suggested removing specific attributes to prevent unnecessary clutter. This included the various variances and the saturation minimum as these attributes are implicit in the vital signs data. Care providers also expressed a desire for enhanced visual cues, proposing that bradycardias, tachycardias, and saturation drops be highlighted in red for quick anomaly identification.

Patient Information The contextual overview of a patient includes measurements relevant to the assessment of a sepsis case (Figure H2). It was requested to incorporate the time elapsed since a measurements was taken. Participants highlighted the importance of temporal context, noting that outdated measures might no longer be relevant.

Abnormal values were highlighted in red, this caused a misconception among care providers who mistakenly believed that values highlighted in red contributed directly to the prediction: *"This is very convenient. I can easily see that the score is 75 because of the skin color of the child and the heart frequency variance."*. However, skin color was not one features used by the prediction model. This highlights the

significance of clear visual cues in aiding their interpretation of predictive elements.

Feature Importance Chart Care providers expressed varied opinions regarding the comprehensibility and utility of the feature importance chart. Doctors found these charts helpful, as they aided in understanding the features utilized by the prediction model and provided insights into the decision-making process. Despite acknowledging its usefulness, doctors suggested that the feature value contribution chart should not have a permanent location on the interface. They also felt that presenting raw feature and effect values, or a reference point, added unnecessary complexity, particularly as the raw values for calculated features like variance were not essential to the diagnosis of sepsis by the doctor. Simplifying the chart to display only the magnitude and direction of the feature was deemed more straightforward.

Both domain experts and usability specialists recommended that the feature value contribution chart be displayed when the user interacts with a specific prediction, either by clicking or hovering over it in the prediction trend chart. Care providers, including doctors, favored this approach as it allowed them to observe how feature value contributions evolved over time. In terms of preferred presentation, nurses leaned towards textual explanations, while doctors (including Neonatal Fellows and physician assistants) showed a preference for visual representations, particularly the tornado plot format (see e.g. Figure H3b). However, nurses, in general, expressed a preference for not having the feature value contribution chart at all, as their focus was on patient care rather than delving into the inner workings of the model. They believed model explanations functioning was more important for doctors.

Risk Scores Both domain and usability experts expressed a preference for displaying risk as categories (e.g. low, medium, high) rather than through numerical values. Healthcare providers incorrectly interpreted the model score as the likelihood of sepsis. However, the model scores are not calibrated and therefore cannot be interpreted as such. Calibration is challenging as the model was trained on a cross-sectional dataset but it is applied as a continuous monitoring system with hourly predictions. Moreover, displaying the exact probability of sepsis to healthcare providers may not be desirable, as the precision is low resulting in low true probabilities even for high-risk alarms. Such low probabilities might not be taken seriously. To make the results more accessible, an approach categorizing outcomes into understandable levels, such as low, medium, and high-risk areas, can be an effective solution [10]. This approach facilitates interpretation, especially for non-technical users, and provides better context for decision-making.

Healthcare providers wanted to know the necessary actions based on the risk level: *"Should I alert a doctor for a medium-risk or high-risk alarm?"*. Nurses specifically expressed the need for clear instructions regarding actions to be taken for each risk score category, asking questions like, *"Do I need to contact a doctor when the risk score is high?"* They emphasized the importance of a conclusive section on the dashboard to facilitate quick decision-making.

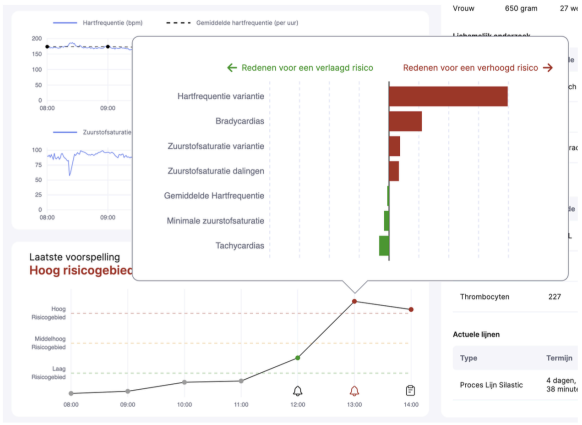
NICU Patient Overview Some domain experts showed a positive response to the unit patient overview component. They thought the information was clear and well structured. Usability experts recommended expanding the patient information when a user clicks on a specific patient in the overview. Additionally, domain experts recommended filtering patients based on the unit to which they belonged, as they were primarily concerned with alarms for patients within their designated units.

Because every alarm signal, regardless of whether it is medium or high, requires patient assessment, the unit patient overview intentionally does not differentiate between low, medium, or high-risk alarms. The system simply uses two colors: red when an alarm is present and and a neutral color if there is no new alarm. Nevertheless, it is possible to see in the trend prediction chart whether the alarm signal is triggered in a low, medium, or high-risk area (Figure H3a). This feature is crucial as it provides insight into the context of the patient, is their status deteriorating rapidly or have they been at high risk for prolonged periods of time? A transition from low to high risk may potentially be more alarming than a shift from medium to high. The boundaries for the risk threshold were carefully determined by the model developers to prevent alarm fatigue [12]. Low-risk alarms are frequent and have high recall at a cost of low precision, whereas high-risk alarms are more rare but have higher precision.

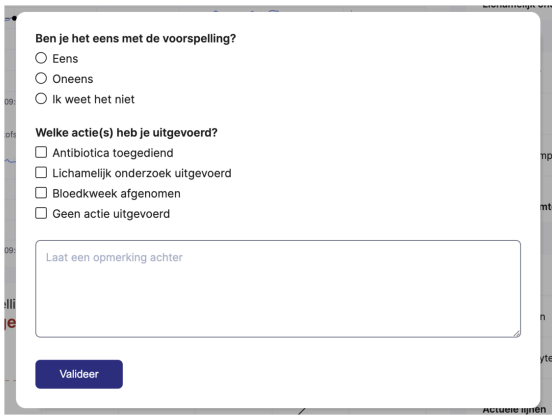
3.4. Implement Phase

Figure 2 shows the high-fidelity prototype that was developed based on the input from the low-fidelity prototype. Pop-up windows are shown in Figure 3: hovering over a risk score provides insights into the feature contributions (Figure 3a) and a form appears when validating an alarm (Figure 3b). A detailed description of the various elements can be found in Appendix I.

Figure 2: Dashboard design of the high-fidelity web application prototype. The dashboard contains an overview of the patients (C1), alarm notification (C2), vital signs (C3), prediction trend (C4) and a summary of the patient information (C5). Note: patient names & dates are not real. Note that text is in Dutch as it was designed for Dutch care providers.



(a) Feature value contribution chart.
Translation of green (red) text: "Reason for decreased (increased) risk"



(b) Alarm validation window.
Translation: "Do you agree with the prediction? Agree, Disagree, I don't know."

"Which action(s) did you perform? Administer antibiotics, Perform physical examination, Draw bloodculture, No action performed"

Figure 3: Pop-up windows in the high-fidelity prototype (original text in Dutch).

Seven caregivers (one male, six female) working at the NICU participated in the evaluation of the high-fidelity prototype. The sample contained a diversity in terms of practical experience and roles, with three physician assistants (P1, P5 and P6), two nurses (P2 and P7), a neonatologist (P3) and a Neonatal Fellow (P4). Due to limitied availability, three participants had also participated in the low-fidelity evaluation. However, there had been a two-month washout period between interviews.

3.4.1. Task-based Think-Aloud Study

The task-based think aloud study consisted of four distinct scenarios, involving a true positive, false positive, true negative and false negative alarm. Results of the study were analysed qualitatively along three themes: trust and reliability, decision making, and comprehensibility and usability.

Trust and Reliability Sub-themes of trust and reliability include the assessment of clinical relevance, alignment with domain knowledge, and perceptions of predictive performance.

Participants relied heavily on contextual patient information to assess predictions. They leaned on this contextual data to assess the clinical relevance of a prediction and the patient's overall risk. This involved discerning deviations, establishing baselines, and excluding alternative explanations. Participants were observed to first analyse the chart with vital signs to identify whether deviations from the normal pattern took place and how severe these were. Then, they used general patient information to determine patient baselines. For instance, one participant (P4) noted: *"In itself, for a baby who is born at 25 weeks and now four weeks old, I don't think a little rumbling saturation is a big deal"*. Providers also relied heavily on actual interventions, such as respiratory support or antibiotics, to determine patient baselines. For instance, it was noted (P2): *"If saturation drops occur while the child is on a ventilator, it is even more worrisome"*. However, providers were cautious in answering how worried they were and whether they agreed with predictions for alarming patients as they indicated a need for more contextual information to assess the clinical relevance. Finally, they stated that before they could trust the prediction, they wanted to be sure that the observed incidents were not due to some other explanation. For instance, P1 remarked: *"I do need to know if it is real [meaning: not due to interventions such as bathing, nursing, feeding]. If I am told: these are real drops then I do have confidence in the prediction.*". They mentioned that measurements could also be unreliable due to subjective measurements, indicating the importance of visiting the patient themselves and acquiring information from other nurses. For example, some participants indicated that the skin colour could also be entered by an inexperienced nurse or could have changed in the meantime. Evaluating the clinical relevance of a prediction appeared a crucial factor influencing care providers' trust in the model. Nevertheless, the available data on the dashboard proved insufficient for a comprehensive assessment.

Care providers mainly relied on the vital signs chart together with the explanations (risk analysis and feature importance chart) to verify if the predictions aligned with their expectations. When predictions aligned, such as in the true positive case, trust in the model increased. P6 noted: *"Because there was really a huge increase in the number of bradycardias that weren't there before, I do understand the higher risk"*. However, the opposite was also true, trust in the model decreased when predictions did not align with clinical knowlege. For instance, P3 noted that the prediction trend increased, while the vital signs stabilized: *"Why is there an alarm? Because here you actually have very stable heart action again"*.

The perceptions of predictive performance also influenced trust. For instance, P2 was reluctant in relying on the model because the model had not yet proven itself: *"If you're going to find in practice that alerting is really an indicator of an infection, then I'll definitely take that on board."*. These findings were confirmed by other participants (P3, P4). Moreover, trust in the predictions was also impacted by limitations of the model. For instance P4 doubted whether they could rely on a model that only used heart frequency and oxygen saturation: *"I do think that heart rate and saturation alone do not cover it, so to speak, for me to decide whether or not to use antibiotics"*. Finally, care providers thought that one prediction per hour was not enough, as mentioned by P4: *"But you can still miss a lot in an hour. It would be best if it were continuous"*. Therefore, providers did not rely on the prediction model to make a decision but used it more as reassurance, to increase confidence or as a second opinion.

Decision Making Similar to findings from the low-fidelity prototype, care providers expressed a need for clarity and meaning regarding the interpretation of predictions. For instance, P3 faced challenges in the interpretation *"how are you going to interpret it? Are you going to say high-risk then by definition you have to start antibiotics."* P3 also struggled to gauge the degree of risk within a given range, stating *"still with high-risk you have a kind of gray area, is this 85% or 99% chance that it could be an infection?"*. This underscores the need for guidance on utilizing the model and interpreting risk predictions.

Predictions can have a noticeable impact on care providers, prompting them to conduct more thorough examinations of the patient or reevaluate their initial assessments. P4 noted that when the prediction diverged from their own assessment, as observed in task 2 (the false positive case) it prompted them to look at the patient more closely. P4 expressed, *"I do find it funny, because I find that this does trigger me to look at the child again. Maybe it really is declining after all."* The contradiction compelled the care provider to reevaluate their initial assessment, even though they were not initially concerned: *"This* [referring to the monitor with vital parameters] *doesn't necessarily worry me, but then again, this* [referring to the latest prediction score] *does a little bit. With the thought that it probably predicts before you would see it, I would like to see the child"*. Also P3 examined the false positive case with extra attention. The participant utilized the model explanations to understand why the algorithm deemed the patient at risk of sepsis, comparing it with vital signs and engaging in reasoning, questioning whether there really was a saturation drop or not.

3.4.2. Usability

Participants found it convenient that relevant patient information was presented in a central overview, rather than having to search for scattered information in the EHR. The centralized overview aided care providers in verifying risk predictions and quickly assessing patients. It was particularly beneficial for those commencing their shifts. Moreover, the centralized overview proved advantageous for presenting cases to a doctor, as noted by P2: *"That certainly means that I get my overall picture clear and can then alert the doctor"*. However, there were concerns as to whether the presented information was actually sufficient and whether adding additional information (e.g., patient history, additional timelines) is feasible, as stated by P3 *"Well, I think you just need more patient information, so to speak, but the question is whether you should put it all in that dashboard."*.

The risk trend and feature importance plot were received positively. Throughout the tasks, participants consistently referred to these, finding it straightforward to comprehend the evolution of risk trends and to discern features influencing the prediction. Although one participant (P1) found the feature importance plot somewhat challenging to interpret. In their effort to understand the explanations, participants not only utilized the feature importance plot but also relied on the vital sign charts.

Some suggestions for improvement were proposed. A majority of participants (P1, P3, P4, and P6), proposed that the vital sign chart would be more valuable for validating predictions if it exclusively displayed validated data, where artefacts due to e.g. feeding and nursing have been removed. Such refinement would streamline the validation process, allowing for more accurate risk assessment without the need for additional checks. Additionally, participants expressed the need to zoom in or out on the vital sign charts to gather patient baselines. This can aid in understanding whether certain patterns, such as saturation drops, are normal for a particular patient. The integration of counts (e.g. bradycardias, tachycardias, saturation drops) next to the vital sign charts was deemed helpful, eliminating the need for manual counting. However, participants stressed the importance of establishing clear thresholds that should align with the criteria utilized in the EHR. Participants also identified areas for improvement in the dashboard design. Some overlooked the demographic patient information, emphasizing the need for greater visual prominence. Moreover, participants advocated the removal of outdated or irrelevant parameters from the dashboard to maintain its relevance and utility.

The average SUS score was 86.25, with excellent scores for four participants (P1, P2, P5 and P6) and marginal ones for two (P3 and P4). Notably, the physician assistants and nurses gave higher scores than the neonatoligst and Neonatal Fellow. Full details are in Appendix E.

4. Discussion and Conclusion

This study designed an XAI interface using the framework from Sedlmair et al. [19] which integrates contextual patient information with the sepsis risk model from van den Berg et al. [12].

RQ1: It was shown that a diverse range of contextual information was needed to interpret the sepsis predictions (RQ1). In agreement with Barda et al. [6] vital sign information proved of pivotal importance, but also other information on patient demographics and interventions such as respiratory support and medication were deemed important. This information was needed to compare predictions against user expectations and to build trust in model predictions. Of lesser importance, but still received positively, were risk trends and feature contributions.

RQ2: Healthcare providers responded positively to the interface with consolidated view (RQ2). Combining risk scores with information on vital parameters and treatment allowed for efficient validation of model predictions. However, information needs remained unfulfilled and it is challenging to present all relevant information in a central overview.

RQ3: Similar to the findings of Matthiesen et al. [10], healthcare providers primarily relied on their established methods to assess patient risk using the model as a confirmation or second opinion (RQ3). In certain cases, when the model prediction did not align with expectation of the practitioner, it caused them to re-evaluate a patient. However, there was skepticism regarding the model performance and the limited number of predictors based solely on heart rate and oxygen saturation also adversely affected model trust. Simultaneously, some participants expressed an interest in similar systems called HeRO which is based solely on heart-rate characteristics [26]. This suggests a potential gap in understanding of such algorithms. Finally, there was a disparity between the intended use of the algorithm and how care providers perceived its application. Caregivers sought definitive actions with risk predictions, such as initiating antibiotic treatment for high-risk patients, contrary to the model's intended purpose of alerting caregivers to patients at elevated risk of sepsis. Such unintended AI influences touch upon an important ethical consideration and a clear gap in research about the intended and applied division of responsibility in clinical decision support systems [27, 28]. It illustrates that users should be made thoroughly familiar with the intended use, through e.g. training, and that the division of responsibility between decision support system and healthcare professional should be explicit.

Limitations: First, the study was conducted away from the bedside where the model ought to operate. Here contextual information, such as patient skin color or the presence of respiratory support, is available by looking at the patient. Requirements regarding contextual information will likely vary in different study settings emphasizing the importance of emulating the context of use as closely as possible. In addition, actual bedside testing could present new challenges related to for instance alarm fatigue from false positives, or arising from working in a context with multiple separate interfaces (bedside monitor, EHR, sepsis tool, etc.). Secondly, the number of participants was limited. Participants partially overlapped between the design and evaluation phases, where some had seen parts of the dashboard earlier, possessing additional knowledge about the model and its limitations. Due to staffing constraints, we had to approach participants, who had previously participated. However, a significant time gap (over two months) between the design and evaluation phases likely resulted in participants forgetting major interface details. Finally, the model under consideration is an imperfect predictor yielding many false positives (low precision). The generalizability of the findings likely do not extent to models that are near-perfect predictors as healthcare professionals might have more trust in these models a-priori and the suggested interventions are more explicit. For instance, a near-perfect sepsis predictor would just prescribe the administration of antibiotics (prescriptive model). However, it is highly unlikely that such model performance is attainable for applications such as a sepsis early warning system in the near future.

To conclude this discussion, we highlight important opportunities and challenges that can guide development activities of clinical decision support tools and future research directions in the context of HCI in the healthcare domain.

Upfront requirement specification through HCI methodology: The study demonstrates the importance of careful gathering of stakeholder requirements and the crucial role of developing interfaces

that present model predictions such that they become actionable for practitioners in clinical practice. This is in line with work on participatory design and human-centered design [29]. For the sepsis use case, a lot of contextual information is required for users to trust and use the model. Leaving out this contextual information likely renders the model ineffective. In order to efficiently develop decision support that aligns with stakeholder expectations, careful design and requirement engineering should be done upfront. However, the design and HCI perspective appears below par in healthcare, where dataset availability often drives initial development decisions rather than careful human-centric requirement engineering, and this likely contributes to the limited adoption of decision support in practice [30].

Relevance of contextual information for monitoring applications: The methodology and findings of this study are relevant for most clinical decision support tools that are not prescriptive. Predictive decision support- or continuous monitoring tools such as the present LOS algorithm require a decision from the end user based on alarms with limited predictive power (e.g. high false alarm rate). In the absence of explicit (prescriptive) protocols, healthcare professionals likely require additional contextual information to assess the likelihood of an event. This is in line with earlier work emphasizing contextualization for explainable AI especially for non-expert users [31].

Integration of multiple decision support aids: A challenge that remains is integrating the information of multiple decision support models such that they can be used effectively by healthcare providers. So far, only limited number of decision support tools are available. The NICU sepsis warning system is just one system, which requires a lot of contextual information to be used effectively. However, in the future many more models will be introduced, possibly with similar demands for additional contextual information. Bespoke interfaces for each model will likely not be feasible, nor desirable so it should be investigated how predictions from all these models together with relevant contextual information can be unified, in the EHR or elsewhere. This should be done in such a way that it aids the clinical workflow and does not result in for instance alarm fatigue or mental overload.

Aligning mental models of users with the intended use: The study underscored instances where contextual information on the dashboard was misinterpreted as a predictor, leading to a misunderstanding of the model's predictions. Subsequent research should focus on refining the presentation of contextual information, employing visual cues to enhance understanding. It became evident that mental models of care providers often did not align with the intended use of the predictive model. Future research in HCI should delve into strategies and design principles that can bridge these misalignments as well as appropriate ways to instruct users on how to use decision support tools. Aligning AI with user's mental models is still an open challenge in many XAI use cases [32].

Communicating uncertainty: Care providers expressed a desire for evidence of the model's performance, yet found technical validation metrics too complex. A critical future research direction involves exploring user-friendly ways to communicate the confidence and reliability of predictive models. This could include the development of intuitive visual indicators, color-coded confidence levels, or the use of plain language to convey the reliability of predictions. Designing for uncertainty is an difficult and unsolved challenge in XAI [33].

In conclusion, this study illustrates the importance of contextual information in interpreting the predictions from a clinical decision support system. It re-establishes the importance of careful requirement engineering and design decisions when developing actionable systems, highlighting the importance of design and HCI specialists and re-emphasizing that a prediction model is only a piece of the puzzle. Awareness of the critical role of design and HCI to unravel how users interact with these systems is paramount to developing safe tools and overcoming adoption challenges. this cannot be overcome.

Acknowledgments

We thank the participants in the surveys for taking the time to participate in this study. In addition, we thank Michael Behrisch and Ruben Peters for useful discussions.

5. Declaration on Generative AI

This work was prepared using Grammarly for grammar and spelling checks. The authors reviewed and edited the final content and take full responsibility for the publication's content.

References

[1] C. L. Cifra, J. W. Custer, J. C. Fackler, A Research Agenda for Diagnostic Excellence in Critical Care Medicine, Critical Care Clinics 38 (2022) 141–157. URL: https://www.sciencedirect.com/science/article/pii/S0749070421000543. doi:10.1016/j.ccc.2021.07.003.

[2] A. L. Beam, I. S. Kohane, Big Data and Machine Learning in Health Care, JAMA 319 (2018) 1317. URL: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.2017.18391. doi:10.1001/jama.2017.18391.

[3] A. Adadi, M. Berrada, Explainable AI for Healthcare: From Black Box to Interpretable Models, in: V. Bhateja, S. C. Satapathy, H. Satori (Eds.), Embedded Systems and Artificial Intelligence, Springer, Singapore, 2020, pp. 327–337. doi:10.1007/978-981-15-0947-6_31.

[4] H. D. Zając, L. Dana, X. Dai, J. F. Carlsen, F. Kensing, T. O. Andersen, Clinician-Facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI, ACM Transactions on Computer-Human Interaction (2023). URL: https://dl.acm.org/doi/10.1145/3582430. doi:10.1145/3582430, publisher: ACMPUB27New York, NY.

[5] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38. URL: https://www.sciencedirect.com/science/article/pii/S0004370218305988. doi:10.1016/j.artint.2018.07.007.

[6] A. J. Barda, C. M. Horvat, H. Hochheiser, 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) 257. URL: https://doi.org/10.1186/s12911-020-01276-x. doi:10.1186/s12911-020-01276-x.

[7] D. Wang, Q. Yang, A. Abdul, B. Y. Lim, Designing Theory-Driven User-Centric Explainable AI, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, ACM, Glasgow Scotland Uk, 2019, pp. 1–15. URL: https://dl.acm.org/doi/10.1145/3290605.3300831. doi:10.1145/3290605.3300831.

[8] Z. Jin, S. Cui, S. Guo, D. Gotz, J. Sun, N. Cao, CarePre: An Intelligent Clinical Decision Assistance System, ACM Transactions on Computing for Healthcare 1 (2020) 1–20. URL: https://dl.acm.org/doi/10.1145/3344258. doi:10.1145/3344258.

[9] S. Sandhu, A. L. Lin, N. Brajer, J. Sperling, W. Ratliff, A. D. Bedoya, S. Balu, C. O'Brien, M. P. Sendak, Integrating a Machine Learning System Into Clinical Workflows: Qualitative Study, Journal of Medical Internet Research 22 (2020) e22421. URL: https://www.jmir.org/2020/11/e22421. doi:10.2196/22421.

[10] S. Matthiesen, S. Z. Diederichsen, M. K. H. Hansen, C. Villumsen, M. C. H. Lassen, P. K. Jacobsen, N. Risum, B. G. Winkel, B. T. Philbert, J. H. Svendsen, T. O. Andersen, Clinician preimplementation perspectives of a decision-support tool for the prediction of cardiac arrhythmia based on machine learning: Near-live feasibility and qualitative study, JMIR Hum. Factors 8 (2021) e26964.

[11] S. A. Coggins, K. Glaser, Updates in Late-Onset Sepsis: Risk Assessment, Therapy and Outcomes, NeoReviews 23 (2022) 738–755. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9675597/. doi:10.1542/neo.23-10-e738.

[12] M. A. M. . van den Berg, O. O. A. G. . Medina, I. I. P. . Loohuis, M. M. . van der Flier, J. J. . Dudink, M. M. J. N. L. . Benders, R. R. T. . Bartels, D. D. C. . Vijlbrief, Development and clinical impact assessment of a machine-learning model for early prediction of late-onset sepsis, Computers in Biology and Medicine 163 (2023) 107156. URL: https://www.sciencedirect.com/science/article/pii/S0010482523006212. doi:10.1016/j.compbiomed.2023.107156.

[13] J. F. Hicks, K. Fairchild, Hero monitoring in the nicu: sepsis detection and beyond, Infant 9 (2013) 187–191.

[14] M. Meeus, C. Beirnaert, L. Mahieu, K. Laukens, P. Meysman, A. Mulder, D. V. Laere, Clinical Decision Support for Improved Neonatal Care: The Development of a Machine Learning Model for the Prediction of Late-onset Sepsis and Necrotizing Enterocolitis, The Journal of Pediatrics 266 (2024). URL: https://www.jpeds.com/article/S0022-3476(23)00733-3/abstract. doi:10.1016/j.jpeds.2023.113869, publisher: Elsevier.

[15] N. Bienefeld, J. M. Boss, R. Lüthy, D. Brodbeck, J. Azzati, M. Blaser, J. Willms, E. Keller, Solving the explainable AI conundrum by bridging clinicians' needs and developers' goals, npj Digital Medicine 6 (2023) 1–7. URL: https://www.nature.com/articles/s41746-023-00837-4. doi:10.1038/s41746-023-00837-4, publisher: Nature Publishing Group.

[16] A. Kaltenhauser, V. Rheinstädter, A. Butz, D. P. Wallach, "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, DIS '20, Association for Computing Machinery, New York, NY, USA, 2020, pp. 1509–1522. URL: https://dl.acm.org/doi/10.1145/3357236.3395436. doi:10.1145/3357236.3395436.

[17] S. M. Gephart, D. A. Tolentino, M. C. Quinn, C. Wyles, Neonatal Intensive Care Workflow Analysis Informing NEC-Zero Clinical Decision Support Design, Computers, informatics, nursing: CIN 41 (2023) 94–101. doi:10.1097/CIN.0000000000000929.

[18] M. Zhang, D. Ehrmann, M. Mazwi, D. Eytan, M. Ghassemi, F. Chevalier, Get To The Point! Problem-Based Curated Data Views To Augment Care For Critically Ill Patients, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22, Association for Computing Machinery, New York, NY, USA, 2022, pp. 1–13. URL: https://dl.acm.org/doi/10.1145/3491102.3501887. doi:10.1145/3491102.3501887.

[19] M. Sedlmair, M. Meyer, T. Munzner, Design Study Methodology: Reflections from the Trenches and the Stacks, IEEE Transactions on Visualization and Computer Graphics 18 (2012) 2431–2440. URL: https://ieeexplore.ieee.org/document/6327248/?arnumber=6327248. doi:10.1109/TVCG.2012.213, conference Name: IEEE Transactions on Visualization and Computer Graphics.

[20] T. Munzner, Visualization Analysis and Design, CRC Press, 2014. Google-Books-ID: dznSBQAAQBAJ.

[21] C. Molnar, Interpretable Machine Learning, Lulu.com, 2020. Google-Books-ID: jBm3DwAAQBAJ.

[22] j. Brooke, SUS: A 'Quick and Dirty' Usability Scale, in: Usability Evaluation In Industry, CRC Press, 1996, pp. 4–7.

[23] J. R. Lewis, The System Usability Scale: Past, Present, and Future, International Journal of Human–Computer Interaction 34 (2018) 577–590. URL: https://doi.org/10.1080/10447318.2018.1455307. doi:10.1080/10447318.2018.1455307, publisher: Taylor & Francis _eprint: https://doi.org/10.1080/10447318.2018.1455307.

[24] J. Lazar, J. H. Feng, H. Hochheiser, Research Methods in Human-Computer Interaction, Morgan Kaufmann, 2017. Google-Books-ID: hbkxDQAAQBAJ.

[25] C. Pribeanu, A Revised Set of Usability Heuristics for the Evaluation of Interactive Systems, Informatica Economica 21 (2017) 31–38. URL: https://ideas.repec.org//a/aes/infoec/v21y2017i3p31-38.html, publisher: Academy of Economic Studies - Bucharest, Romania.

[26] K. D. Fairchild, Predictive monitoring for early detection of sepsis in neonatal ICU patients, Current Opinion in Pediatrics 25 (2013) 172. URL: https://journals.lww.com/co-pediatrics/fulltext/2013/04000/predictive_monitoring_for_early_detection_of.5.aspx. doi:10.1097/MOP.0b013e32835e8fe6.

[27] L. Crompton, The decision-point-dilemma: Yet another problem of responsibility in human-AI interaction, Journal of Responsible Technology 7-8 (2021) 100013. URL: https://www.sciencedirect.com/science/article/pii/S2666659621000068. doi:10.1016/j.jrt.2021.100013.

[28] J. Zerilli, A. Knott, J. Maclaurin, C. Gavaghan, Algorithmic Decision-Making and the Control Problem, Minds and Machines 29 (2019) 555–578. URL: https://doi.org/10.1007/s11023-019-09513-7. doi:10.1007/s11023-019-09513-7.

[29] I. Göttgens, S. Oertelt-Prigione, The application of human-centered design approaches in health research and innovation: a narrative review of current practices, JMIR mHealth and uHealth 9 (2021) e28102.

[30] J. Kernahan, R. Bartels, M. de Reuver, D. Oberski, R. Dobbe, Burying the lead: Adjusting goals to manage functional limitations of ai tools in healthcare, Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society 8 (2025) 1401–1412. URL: https://ojs.aaai.org/index.php/AIES/article/view/36640. doi:10.1609/aies.v8i2.36640.

[31] C. Bove, J. Aigrain, M.-J. Lesot, C. Tijus, M. Detyniecki, Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users, in: Proceedings of the 27th international conference on intelligent user interfaces, 2022, pp. 807–819.

[32] M. Merry, P. Riddle, J. Warren, A mental models approach for defining explainable artificial intelligence, BMC Medical Informatics and Decision Making 21 (2021) 344.

[33] S. Cao, A. Liu, C.-M. Huang, Designing for appropriate reliance: The roles of ai uncertainty presentation, initial user decision, and user demographics in ai-assisted decision-making, Proceedings of the ACM on Human-Computer Interaction 8 (2024) 1–32.

[34] K. Koffka, Principles of Gestalt Psychology, Psychology Press, 1999. Google-Books-ID: cLnqI3dvi4kC.

[35] T. A. J. Schoonderwoerd, W. Jorritsma, M. A. Neerincx, K. van den Bosch, Human-centered XAI: Developing design patterns for explanations of clinical decision support systems, International Journal of Human-Computer Studies 154 (2021) 102684. URL: https://www.sciencedirect.com/science/article/pii/S1071581921001026. doi:10.1016/j.ijhcs.2021.102684.

[36] R. M. Monarch, Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI, Simon and Schuster, 2021.

A. Informed Consent

A.1. Purpose of this Study

As part of a master's program in human-computer interaction at Utrecht University, this study is being conducted during an internship at UMCU. The primary objective of this study is to determine the requirements of critical care professionals in the NICU for the design of a new interface for a sepsis risk prediction model.

Human-computer interaction is a multidisciplinary field that combines information, computer science, and psychology to implement technology in realworld situations. One of the major research areas in this field is human-centered machine learning, which aims to ensure that AI decisions are made with a human-centric approach. As AI continues to become more prevalent, algorithms are increasingly being used for medical diagnoses, making it crucial to involve healthcare professionals in the design of such systems.

The NICU at WKZ has developed an AI model that can predict the risk of sepsis in babies before symptoms appear. This model has the potential to aid critical care professionals in intervening earlier, thereby reducing suffering. However, the algorithm currently lacks an interface, which could be critical to the success of the model. A poorly designed interface could hinder the effectiveness of the model and prevent its adoption by healthcare professionals. Moreover, it is essential for care professionals to understand how the model works so that they can decide when to rely on the model and when to rely on their expertise.

- The researcher has explained the purpose of the research to me.
- I have had an opportunity to ask questions about the study.

A.2. Freedom to Withdraw

Your participation in this study is voluntary.

- You can refuse to take part at any time.
- You can take a break at any time.
- You can ask questions at any time.
- I understand that I can leave at any time without giving a reason.

A.3. Privacy and Confidentiality

The interview will be recorded. After the recording is transcribed, it will be deleted. No one else will get to hear the recordings. This means you will not be identifiable, and your comments will be confidential. We may publish research reports that include your comments. The data used in these reports will be anonymous.

- I understand that my voice will be recorded.
- I understand that my comments are confidential.

A.4. Your Agreement

To take part in the research, please sign this form showing that you consent to us collecting these data.

B. Discover: Interview Outline

B.1. General

- Can you tell me about the last time a child was suspected of sepsis? Follow-up questions:
  - How did you assess the seriousness of the situation? Do you always take the same approach to assessing such situations?
  - Which moment is decisive for you to administer antibiotics?
  - What challenges do you face in diagnosing sepsis in the ICU?

B.2. Analysing Information

- What strategies or techniques do you use to research and analyze the available information? (e.g. clinical assessment (vital signs), history, lab, consultation, guidelines)

B.3. Decision-Making

- How does the available time affect your decision-making process?
  - Is your information need and how you analyze information different in an acute situation?

B.4. Prediction Algorithm

- How do you feel about using AI-based systems in the NICU?
- How do you see an ideal interface for an algorithm that predicts the risk of sepsis every hour? Which features would be most useful to you?
- What additional data or contextual information would be valuable to have in addition to the risk score to support your decision making process?
- If the AI-system makes a false positive or false negative prediction, how do you want the interface to let you know about this?
- How could the predictive algorithm support you in diagnosing sepsis?

C. Discover: Interview Outline - Information Sharing

C.1. General

- Can you tell me about the last time a child was suspected of sepsis?

C.2. Sharing Information

- Can you tell me how you currently share information with other healthcare providers?
- Is there information that you think is crucial for other healthcare providers, but that might be missed in the current process?
- Are there any problems or challenges you encounter when sharing information with other healthcare providers?

C.3. Using Information

- Can you tell me how you currently receive information from other healthcare providers?
- Which information from other healthcare providers is crucial to you?
- Is there information from other healthcare providers that you think is crucial, but don't have access to?
- Are there any challenges you face in getting information from other healthcare providers?

C.4. Collaboration

- What challenges do you encounter when collaborating with other healthcare providers?

D. Design: Interview Outline Domain-Experts

- Can you explain why the model made this prediction?
- What information has led you to trust or not trust the prediction?
- Is there any information missing that could make understanding this predictions easier?
- Is there any information that you don't find useful or doesn't help in understanding this prediction?
- What information do you find most important to see at a glance?
- Do you think the current level of detail is enough, or would you like even more detailed information?
- What visualisations do you find the easiest to understand?
- What do you think about how the information is organized on the screen? For instance, all information in one screen or do you prefer it to be in different sections?
- Can you do everything with the information that you want to do or are there other actions you would like to perform? For example, see more details, zoom in, compare?
- Is there anything else you would like to change about these visualisations?

E. System Usability Scale (SUS)

E.1. Questionnaire

Comprising 10 questions, participants express their agreement with each question on a Likert scale ranging from (1) strongly disagree to (5) strongly agree. The scoring methodology involves subtracting 1 from odd statements and 5 from even ones. The cumulative result is then multiplied by 2.5, resulting in a total of 100 points possible.

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

E.2. Results

Table E1: SUS scores

Question	P1	P2	P3	P4	P5	P6
1	Strongly Agree	Strongly Agree	Neutral	Strongly Agree	Strongly Agree	Strongly Agree
2	Strongly Disagree	Strongly Disagree	Disagree	Strongly Disagree	Strongly Disagree	Strongly Disagree
3	Strongly Agree	Strongly Agree	Agree	Neutral	Strongly Agree	Strongly Agree
4	Strongly Disagree	Strongly Disagree	Disagree	Disagree	Agree	Strongly Disagree
5	Strongly Agree	Agree	Disagree	Agree	Strongly Agree	Strongly Agree
6	Strongly Disagree	Strongly Disagree	Neutral	Disagree	Strongly Disagree	Strongly Disagree
7	Strongly Agree	Strongly Agree	Agree	Strongly Agree	Strongly Agree	Strongly Agree
8	Strongly Disagree	Strongly Disagree	Neutral	Neutral	Strongly Disagree	Strongly Disagree
9	Strongly Agree	Strongly Agree	Neutral	Neutral	Agree	Strongly Agree
10	Strongly Disagree	Strongly Disagree	Disagree	Agree	Strongly Disagree	Strongly Disagree
Score	**100**	**97.5**	**60**	**70**	**90**	**100**

F. Evaluation: Task-Based Think-Aloud Protocol

F.1. Demographic Questions

- Age
- Domain-expertise
- Years of experience

F.2. TASKS [repeat for 4 tasks]

Task 1: An alarm was raised at 13:00 for a patient at high risk of sepsis. Are you concerned and how high do you estimate the risk?

If not already mentioned by the participant: Can you describe to me in your own words how you have used the information on the dashboard to come to a decision?

- What actions did you perform? Why those?
- Did you agree with the prediction? Why/why not?
- How sure are you of your decision?
- Could you rank the following components by how much you rely on them for estimating risk?
- Do you think the dashboard contains enough information?
- How difficult did you find the task? (cognitive load)

F.3. Questions After Tasks are Completed

- Would you have visualized it the same way or differently?
- Do you think the dashboard would improve your decision-making? Or do you think existing dashboards such as Metavision provide enough support?

G. Requirements

The identified requirements are specified in Tables G1, G2, G3, G4.

Table G1: Requirements “explanation”

ID	Requirement	Participant	Priority
1.1	The user should be able to see what features contributed to the prediction	P1,P2,P3	Must have
1.2	The user should be able to see the latest model prediction trend	P2, L	Must have
1.3	The user should be able to adjust the timeline of the model predictions trend	P2, L	Should have

Table G2: Requirements ”patient information”

ID	Requirement	Participant	Priority
2.1	The user should be able to see actual interventions (lines, ventilation, antibiotics)	P6, L	Must have
2.2	The user should be able to see the latest results from physical examination (skin colour and activity)	P1, P2, P3, P5, P6, P7	Must have
2.3	The user should be able to see the latest infection parameters. (CRP, Leukocytes, Thrombocytes)	P1, P2	Must have
2.4	The user should be able to see the vital signs over the last 6 hours (heart frequency, mean heart frequency and saturation)	P1, P2, P3, P6, P7, L	Must have
2.5	The user should be able to see the latest result of the blood culture (datetime and type of bacteria)	P2, P3, P6	Must have
2.6	The user should be able to see general patient information (sex, latest weight, gestational age, postnatal age)	L	Must have
2.7	The user should be able to hover over the aforementioned datapoints to see the three latest values	L	Should have

Table G3: Requirements “model, feedback and collaboration”

ID	Requirement	Participant	Priority
3.1	The user should be able to indicate whether they agreed with the prediction	L	Should have
3.2	The user should be able to indicate what actions they performed	L	Should have

Table G4: Requirements “alarm notification”

ID	Requirement	Participant	Priority
4.1	The user should be warned when the predicted risk exceeds the threshold	P1, P3	Must have

ID	Requirement	Participant	Priority
4.2	The user should be able to see when an alarm was generated	Feedback from design phase	Must have
4.3	The user should be able to validate an alarm	Feedback from design phase	Should have
4.4	The user should be able to mute an alarm	Feedback from design phase	Should have

H. Low-Fidelity Prototypes

H.1. Vital signs

Two variants for representing the vital signs were designed: a line chart featuring multiple attributes (Figure H1a) and a variant with a separate line chart per attributes (Figure H1b). The attributes included all features that were used by the model with the exception of countable predictors (e.g. bradycardia): heart frequency, saturation, mean heart frequency, heart frequency variance, saturation variance, and minimum saturation. Munzner [20] advocates for the use of a line chart to illustrate trends when the data involves at least one quantitative attribute (e.g., heart rate frequency) and one ordered attribute (e.g. time). However, when confronted with numerous attributes, the chart can become challenging to interpret, necessitating separate line charts. Separate line charts have the disadvantage that it is more challenging to compare different attributes.

H.2. Patient Information

Designing contextual patient information was deemed straightforward as only the latest values needed to be presented. Guided by the proximity design principle from Gestalt psychology, groups of similar elements were created to enhance interpretability [34]. Furthermore, a color-coded system was implemented to signify the status of values: red indicated alarming values, green represented normal values, and neutral colors were assigned when no specific meaning was attributed.

H.3. Feature Importance Charts

In designing the feature importance charts, three relevant design solutions were found: tornado plots (Figure H3b), textual representation (Figure H3c) and a hybrid textual/visual approach in an ordered table (Figure H3d).

In tornado plots factors contributing to an increased risk score are positioned on the right side while those diminishing the score are located on the left side. Tornado plots were selected over alternative visualizations such as force plots, which were find hard to interpret for individuals lacking AI expertise [6]. This design aligns with the XAI and visualisation literature, highlighting the ease of interpretation and suitability for comparing diverse attributes [20]. Similarly, scatter plots were dismissed due to incongruities with the underlying data structure; they necessitate two quantitative attributes, while feature importances inherently involve a single categorical and quantitative attribute.

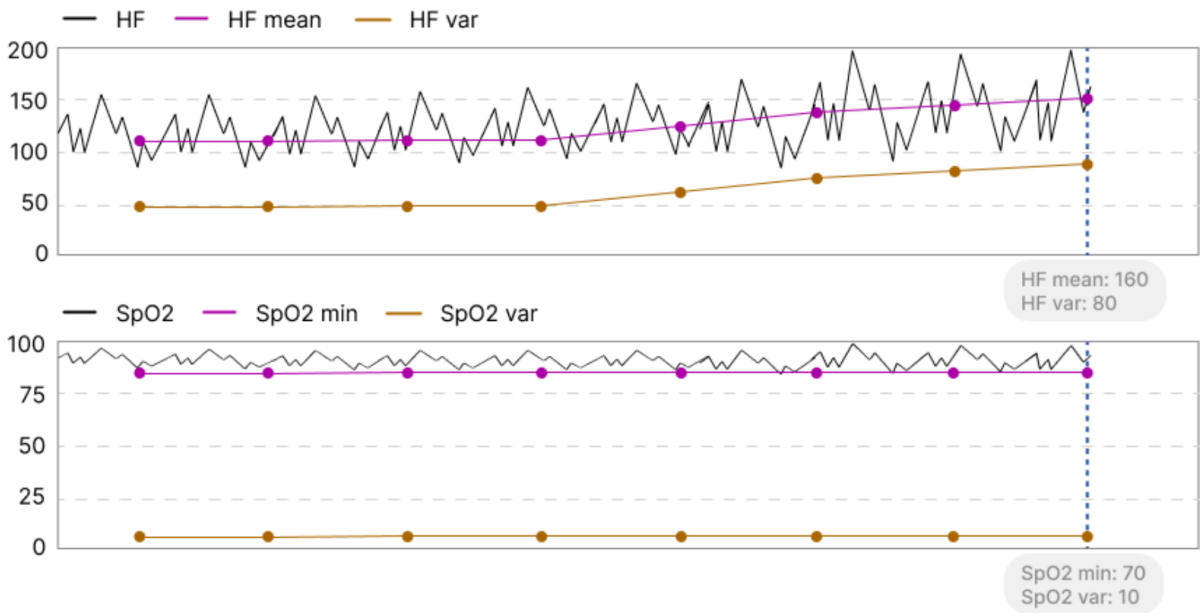
The preference for modality to presenting explanations can diverge between various users [35]. For instance, one group might favor textual over visual presenation. For this reason, textual explanations (Figure H3c) were also included as a design option. Finally, a hybrid solution was also developed, presenting a combination of textual and visual explanations in a tabular format (Figure H3d). Although this approach is visually less telling, it facilitates rapid understanding, especially when dealing with a limited number of features. Tables offer the added benefit of allowing items to be sorted, enhancing the user's ability to focus on specific aspects of interest.

H.4. Risk Scores

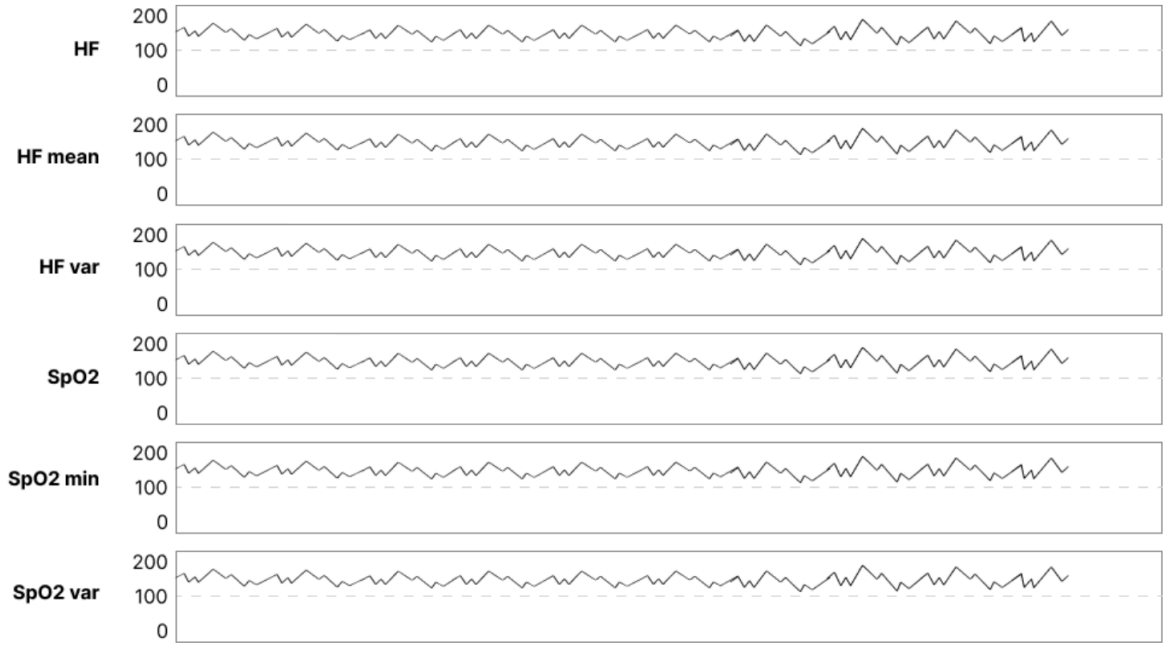
Two low-fidelity prototypes were created to present patient risk. One with categorical labels (Figure H4a) and another with numerical values (Figure H4b). Despite the prevalent use of numerical values in existing XAI interfaces categorical representations were found easier to interpret [15, 6, 9, 10].

H.5. NICU Unit Patient Overview

The risk scores were used to create an overview of all patients, with the title specifying the relevant NICU unit (Figure H5a). During the interviews with care providers, as detailed in the subsequent section, care providers indicated a need to validate and mute alarms (Figure H5b). Nurses, for instance, should



(a) Multi-attribute line chart



(b) Individual line charts

Figure H1: Vital signs. Top: multi-attribute line chart aggregating all heart rate/oxygen saturation information in a single graph. Bottom: individual graphs for each component.

have the capability to validate alarms, signifying that they have attended the patient. Moreover, doctors are granted the authority to mute alarms, facilitating an additional layer of scrutiny.

H.6. Feedback

Feedback options should be shown as binary choices wherever possible. Monarch [36] state that people are more reliable when asked to rank two items rather than judging a problem on a continuous scale. It also is much quicker to hit a button than to drag a slider. Easy-to-use interfaces are particularly important in the ICU to reduce interaction time and cognitive load, especially as these tasks are not directly linked to providing care.

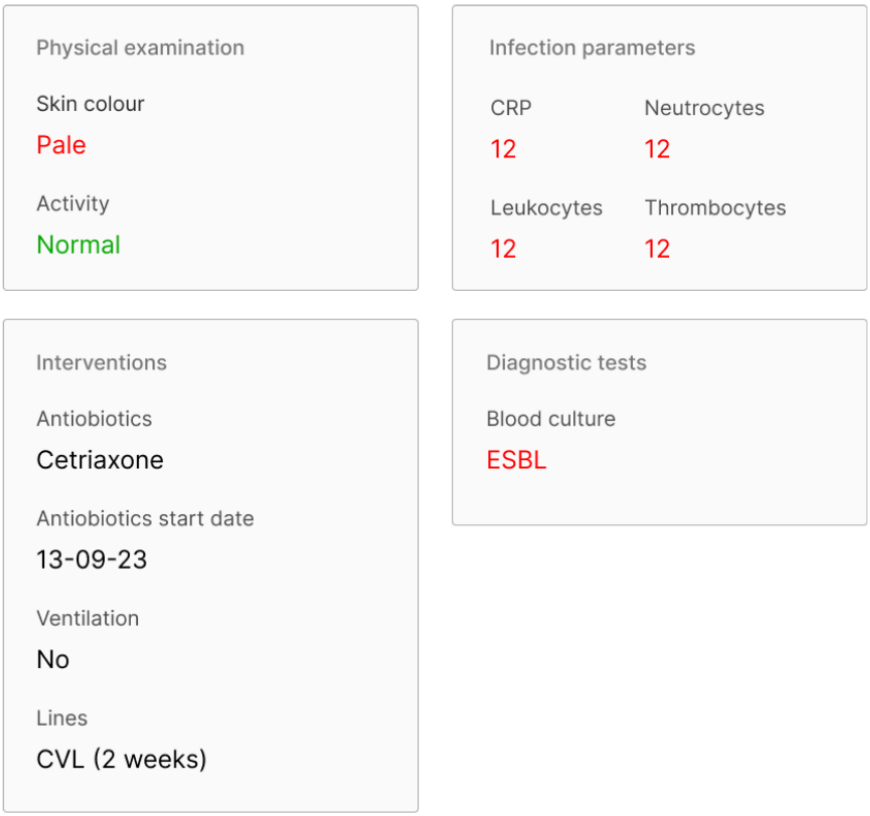
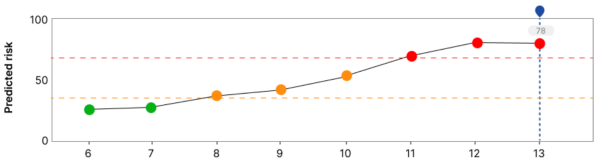
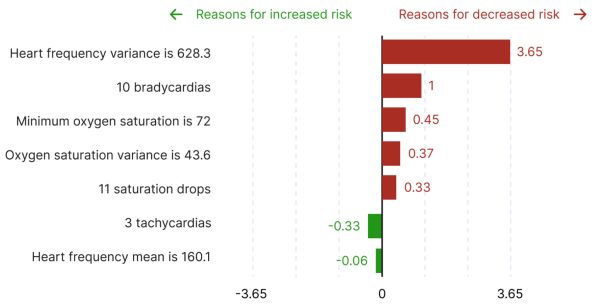


Figure H2: Contextual patient information. The overview contains relevant information not considered by the model, such as skin color and activity level (top left), whether the patient is on antibiotics or being ventilated (bottom left), and laboratory parameters (right).



(a) Risk prediction trend



(b) Tornado plot with feature contributions

A patient's sepsis risk score is influenced by key factors, including a higher heart rate mean (30% contribution) and increased heart rate variance (17%), both of which significantly raise the risk. Bradycardia, indicated by an increasing trend, contributes positively (12%). Conversely, a decrease in the minimum oxygen saturation level has a substantial negative impact (-16%) on the risk score.

(c) Textual description of feature contributions

Number	Feature	Raw value	Trend	Contribution
1	HF variance	628.3	Increase	3.65
2	Bradycardias	10	Increase	1
3	SpO2 min	72	Increase	0.45
4	SpO2 var	43.6	Increase	0.37
5	SpO2 drops	11	Increase	0.33
6	Tachycardias	3	Decrease	-0.33
7	HF mean	160.1	Decrease	-0.06

(d) Table with hybrid (textual and visual) presentation of feature contributions

Figure H3: Sketches providing information about model predictions.

(a) Textual description of feature contributions

(b) Feature contribution table

Figure H4: Risk score representation. On the left, a three-tier risk score is used (low, medium, high), on the right, the model prediction is directly used. Users preferred the representation on the left.

(a) Overview of all patients with their risk status.

(b) Patient overview with the option to validate, mute or reset alarms for each patient.

Figure H5: Overview of patients at the ward

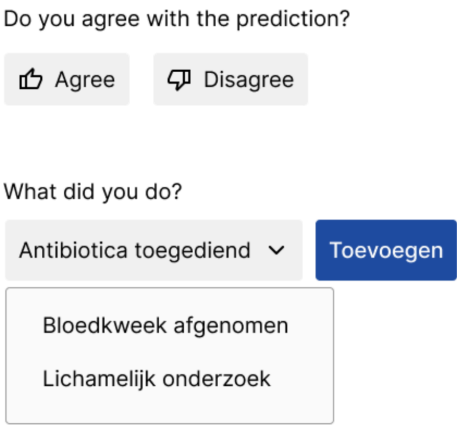


Figure H6: User feedback interface. It asks whether the healthcare professional agrees with the prediction (top) and what action was performed (bottom).

I. High-Fidelity Prototype

Figure 2 in the main text shows the high-fidelity prototype. The dashboard contains an overview of the patients (C1), alarm notification (C2), vital signs (C3), prediction trend (C4) and a summary of the patient information (C5). Here we provide additional details on the elements C1-C5.

I.1. C1: NICU Unit Overview

The user is provided with an overview of all patients associated with the specific unit presented in the dashboard. This ensures that care providers are exclusively notified about patients within their designated unit. However, physicians have access to an overview encompassing patients from all units. When an alarm is triggered for a patient, the vertical bar on the patient tile turns red, accompanied by the display of a red alarm symbol. If the alarm is not validated, a message indicates that the patient is not validated. Conversely, when the alarm is validated, a message confirms the validation status. For patients without alarms, the vertical bar on the left side of the patient tile remains gray, and no alarm icon is visible. The color is either red or gray. No distinction in color is made based on the level of risk because healthcare providers need to assess the patient for every alarm.

I.2. C2: Alarm Notification

When an alarm is triggered, the system records the time of the alarm. Users have the option to validate the alarm by selecting the validation button, which opens a validation form prompting the user to confirm their agreement with the prediction and report the actions taken (maint text Figure 3a). Once validated, the system displays the time at which the alarm was validated. The actions performed can be viewed by other care providers by hovering over the document symbol in the trend prediction chart. Only a physician has the authority to deactivate an alarm. This precautionary measure ensures a clear division of responsibilities regarding the decision-making process. Deactivating an alarm carries potential consequences, and it is crucial to do so under the supervision of a doctor who can assess the situation comprehensively.

I.3. C3: Vital Signs

Users have the ability to review the heart frequency and oxygen saturation indicators spanning the past six hours. By hovering over the graph, users can access precise timestamps and corresponding values corresponding to the mouse pointer location. To streamline the charts and enhance clarity, metrics such as *heart frequency variance*, *oxygen saturation variance* and *oxygen saturation minimum* are not included, as care providers deemed them redundant and a source of unnecessary visual clutter. The charts proved highly beneficial for care providers, enabling them to evaluate the frequency, duration, and intensity of occurrences related to bradycardias, tachycardias, and saturation drops. As a valuable addition, care providers proposed the inclusion of a counter on the dashboard to display the number of tachycardias, bradycardias, and saturation drops, which is presented in the patient information table.

I.4. C4: Prediction Trend Chart

The prediction trend chart empowers care providers to observe the evolution of prediction values. They can customize the time window to assess the development over an extended period of time. Moreover, alarms at the bottom of the chart enable care providers to pinpoint predictions that triggered an alarm, with an active alarm highlighted in red. The universally recognized bell icon ensures instant recognition for care providers. Additionally, care providers can track when an alarm was validated and review associated actions by hovering over the document symbol. For a detailed understanding of how a prediction was generated, a pop-up screen displays the feature value contributions when a user hovers over them (main text Figure 3a).

I.5. C5: Patient Information

Care providers expressed a preference to have all patient information in one central location to facilitate swift decision-making. During the interviews, the response to this was largely positive, with care providers commending the clarity and organization of the information. However, there were some minor points identified for improvement. One crucial addition was the inclusion of timestamps to indicate the recency of the latest metrics. Care providers emphasized the importance of knowing whether the infection parameters were measured recently or a month ago, as it significantly influences decision making. To address this, timestamps were incorporated into a table format, ensuring the information remains easily readable.

Additionally, there were suggestions for additions and removals. Care providers noted the absence of temperature as a parameter and recommended its inclusion in the dashboard. On the other hand, neutrocytes were deemed expendable and were subsequently removed. There was a common misconception regarding the red and green parameter values, with care providers associating them with contributing to predictions. To rectify this, the color coding was removed, and alarming values were highlighted with bold font for a distinctive visual cue. This adjustment aimed to eliminate any ambiguity and enhance the overall interpretability of the information.