# Design of an Explainable Interface for a Sepsis Prediction Algorithm

N.T. Helmantel (6736092)

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Master of Science
in
Human Computer Interaction
under the supervision of
Hanna Hauptmann, Utrecht University
Michael Behrisch, Utrecht University
Daniel Vijlbrief, UMC Utrecht
Richard Bartels, UMC Utrecht

Utrecht University

Department of Computing Science

January, 2024

# Abstract

Machine learning (ML)-based tools hold great promise in clinical practice, as evidenced by lab-based studies. However, it remains particularly challenging for these tools to be officially approved. Insights from Human-Computer Interaction (HCI) show that many face critical user interface issues. A notable issue is the absence of contextual patient information alongside ML predictions. Therefore, 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 (NICU) at the Wilhelmina Children's Hospital. The interface aims to present contextual patient information in a centralized view, offering support to healthcare providers in decision-making and facilitating the interpretation and validation of ML predictions. This research contributes to an overarching HCI topic on designing trustful and purposeful interfaces for ML-based systems in clinical practice. The research starts with the identification of healthcare providers' needs, which were subsequently translated into a set of potential requirements. With this knowledge, a potential solution was created through iterative design processes, developed in collaboration with healthcare providers. To measure the effectiveness of our design, a task-based think-aloud study was conducted, allowing healthcare providers to interact with and provide feedback on the interface. The results from this study highlights the importance of contextual information in interpreting ML predictions. Moreover, presenting all relevant contextual information in a central overview supported decision-making and validating ML predictions, potentially fostering an increased trust in the ML model. The insights gained from healthcare providers lay a solid foundation for future research. Subsequent research can delve deeper into refining the proposed interface, leading to advancements in explainable interface design in clinical practice.

# Contents

# List of Abbreviations

| Abbreviation | Description |
|---|---|
| AI | Artificial Intelligence |
| CRP | C-reactive protein |
| EHR | Electronic Health Record |
| HCI | Human-Computer Interaction |
| HITL | Human-in-the-loop |
| ICU | Intensive Care Unit |
| LOS | Late-Onset Sepsis |
| ML | Machine Learning |
| NICU | Neonatal Intensive Care Unit |
| PICU | Pediatric Intensive Care Unit |
| SHAP | SHaply Additive exPlanations |
| UX | User Experience |
| VA | Visual Analytics |
| WKZ | Wilhemina Kinderziekenhuis |
| XAI | Explainable Artificial Intelligence |

# 1 Introduction

Diagnosing and treating medical conditions within the intensive care unit poses an intricate challenge, complicating the delivery of appropriate care. Caregivers have to analyse and piece together large amounts of information amidst cognitive 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's in patients' conditions.

The combination of ML and the vast amounts of data stored in the electronic health record presents a possible potential for addressing these challenges. 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, it has appeared that many applications fail in practice [2]. A major factor contributing to this issue is the inadequate interpretability of ML models [3] [4]. This concern becomes especially pronounced in healthcare, where decisions have a major impact on patients' well-being. When a healthcare provider is unable to grasp the process that led to a prediction, then it becomes difficult to determine the prediction's reliability. This issue becomes even more critical in instances of errors or unforeseen outcomes.

By far the majority of research aimed at enhancing the interpretability of ML models, known as 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]. For instance, if an explainable model presents a complex representation of the feature space, while an end user lacks experience in machine learning, it becomes futile. This underscores the necessity of user-centric XAI research, especially as an increasing number of machine learning models are being implemented in practice.

HCI researchers have developed frameworks and design 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 underscored the importance of contextual information to comprehend and trust predictions. To illustrate, as elucidated in Jin. et al.'s study [8], historical events are needed as evidence to determine whether a prediction is reliable. In a similar vein, nurses who received sepsis alerts felt uncomfortable making decisions because of the lack of sufficient patient information [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, almost no research has been done on it.

This thesis contributes to an HCI-related issue in which research is conducted on understanding and designing the interaction between humans and intelligent systems. A crucial component of this is the development of user-friendly XAI interfaces to make AI decisions understandable for end users. By investigating both the information needs and presentation preferences of end users, this study aims to contribute to the design of an effective XAI interface.

## 1.1 Sepsis prediction algorithm

Sepsis is an major cause of neonatal morbidity and mortality in the first week of life in the newborn. Every year, 25% of the 1500 premature babies have to deal with it in the Netherlands. From this group, 75 vulnerable infants die [11]. Diagnoses is complex, which leads to sepsis often being diagnosed too late, resulting in serious consequences with death as a worst case scenario. The blood pressure lowers quickly, which leads to the failure of vital organs, such as the heart, brain, and liver. Survivors may suffer from severe and painful residual damage. For this reason, it is important that patients receive effective treatment in a timely manner.

The NICU at the Wilhelmina Children's Hospital developed a novel sepsis prediction algorithm that promises to predict Late-Onset Sepsis (LOS) before symptoms occur, allowing healthcare providers to start treatment in a timely manner. The algorithm is trained on a large dataset with patient records, using heart rate and oxygen saturation as predictors. Currently, the algorithm is being tested under the hood, which means that care providers are not able to act on the information provided by the system, but only to evaluate the accuracy of its predictions in a real-life setting.

The predictions of the model can have a major impact or even do harm when misinterpreted on the patient's condition. Because of this, healthcare providers need sufficient information about both the model and the patient to interpret and validate predictions. If the model predicts a high risk score then care providers need to know what the prediction is based on so it can be validated with clinical information. For example, if the model predicts an increased risk based on an elevated heart rate and the child has just been administered medication then that may be a consideration to wait a while before treating. This is consistent with related studies [6] [9] [12] who found that healthcare providers need contextual patient information, such as vital parameters, laboratory results and patient history to interpret risk scores.

As the condition of the patient can deteriorate rapidly, it is crucial that predictions can be easily interpreted and decisions can be made quickly [13]. Currently, patient information is scattered across various sources, requiring health care providers to search for information in many different places and to do a lot of clicking [14] [15]. As observed in the research conducted by Gephart et al. [15], which surveyed healthcare providers' requirements for a prediction model tar-

geting Necrotizing Entercolitis (NEC), it becomes evident that these providers necessitate a centralised location for accessing aggregated data relevant to the NEC risk score. One healthcare provider noted: *"it seems like it would be very simple to create just a section where all of that information (e.g., NEC risk scoring, clinical signs, feeding information, exposure to preventive treatments) flowed over into so it would give a much quicker snapshot."* Several studies propose that displaying information through an interactive timeline with various synchronized views promoted interpretation [16] [14]. Despite the documented association between poor information presentation and model interpretation [4], little research has been done on it.

## 1.2 Research questions

In conclusion, this thesis aims to bridge the gap between XAI and dashboard design by investigating how contextual patient information should be incorporated into an XAI interface for a sepsis prediction model in the neonatal intensive care unit (RQ1). This study will outline the specific information requirements of healthcare providers (RQ1.1) and explore the efficient presentation of this data by using principles and theories of information visualization (RQ1.2). By combining insights from XAI, dashboard design and healthcare providers needs, this research aims to create an XAI interface that enhances the interpretability of ML models (RQ1.3), ultimately supporting informed decision-making in healthcare.

RQ1: What does an XAI interface with contextual information in one consolidated overview look like to support healthcare providers in interpreting sepsis predictions?

- RQ1.1: What specific contextual information do healthcare professionals require to interpret sepsis predictions?
- RQ1.2: How can contextual information be best presented in an XAI interface to facilitate the interpretation of sepsis predictions?
- RQ1.3: What is the impact of a dashboard with contextual information on trust, reliance and decision-making for healthcare professionals?

# 2 Related work

## 2.1 Artificial intelligence

This subsection offers a concise overview of artificial intelligence, primarily drawing from the content provided in Tom Taulli's book, "Artificial Intelligence Basics" [17]. AI refers to the development of computer systems that can perform tasks that typically require human intelligence. Examples of such tasks are: learning from data to make predictions, understanding and interpreting human language, and interpreting visual information.

### 2.1.1 Machine learning

Machine learning is a subfield of AI that learns from data using statistical models to make predictions without being explicitly programmed. There are two main types of machine learning algorithms: supervised and unsupervised machine learning.

Supervised machine learning algorithms use labeled data. In this approach, input features are matched to their corresponding target labels. The objective is to learn a mapping function that can predict the output given new, unseen inputs. During training, the input-output pairs are presented to the model, optimising its parameters until the discrepancy between the predicted outputs and true labels are minimised. This knowledge is generalised to make predictions on unseen data. There are two main types of supervised learning algorithms. The first is classification, in which the algorithm divides the dataset in common labels. In this case, the target variable is a discrete, categorical value. Examples of these algorithms include Naive Bayes Classifiers and k-nearest neighbors. The other type are regression algorithms, which finds continuous patterns in the data. Examples of these are, linear regressions, ensemble modelling, and decision trees.

Unsupervised learning deals with unlabeled data. This means that there are no corresponding target labels to the input features. The goal of the algorithm is to detect patterns without prior knowledge of the outputs. By identifying similarities or differences between data points, they can categorize or group the data into clusters. Common techniques are clustering algorithms, such as K-means.

Machine learning is a powerful tool that can be used to support humans in their daily lives. Tasks can be automated so that humans can focus on creative and complex work, which may have both economical (e.g., increased productivity) and societal (e.g., increased mental health) benefits. For example, incoming emails can be automatically classified into different categories, such as incoming requests in a call center.

Although machine learning performs well in specific domains and for simple tasks, they do not possess general intelligence that is comparable to human cognition such as abstract reasoning, creativity, and empathy [18]. It depends on large volumes of data and is not able to generalise beyond specific patterns

observed during training.

## 2.2 Human-in-the-loop machine learning

The content on human-in-the-loop machine learning and annotation interfaces primarily relies on information sourced from Robert Monarch's book "Human-in-the-loop Machine Learning" [19]. The synergy of human expertise and machine learning can lead to powerful solutions. On the one hand, machine learning can recognise complex patterns in large amounts of information that is not feasible to comprehend with the human brain. On the other hand, the human brain possesses general intelligence, domain knowledge, creativity and critical thinking skills that can complement data-driven models. Humans can provide valuable insights, but also interpret results and guide decision-making, which can lead to well-informed decisions.

Human-in-the-loop machine learning (HITL ML) not only enables models to solve more complicated tasks, but also makes them insightful for its users. The methods that are used to enable users to provide the system with feedback also allows them to gain insights into the inner working of the system. For example, asking users what features are most appropriate for a dataset also teaches them what features are used to make predictions. Allowing an AI system to edit, improve and repair when it has made mistakes leads to seamless collaboration between humans and the algorithm [20]. This guideline was empirically augmented in a follow-up study [21] by the same authors in which it was confirmed that allowing users to provide feedback increases trust and interpretability.

### 2.2.1 Human experts in the ML process

HITL ML is a branch of machine learning that focuses on the improvement of machine learning models through interaction and feedback with the system. [22] identified several stages to which HITL ML can be applied:

- **Data producing:** Human experts can assist in labeling raw data so that it becomes training data. For example, they can manually label the sentiment of a tweet as positive, negative or neutral. More complicated ways of annotating data consists of transcribing audio or generating texts.
- **Data preprocessing:** Humans can assist in data preprocessing or data cleaning to detect and fix errors in the dataset. For example, they can deal with outliers, inconsistent data, and missing values.
- **Feature selection:** Humans can cooperate in the process of selecting and generating relevant features based on their domain knowledge. This is particularly relevant when there is too little training data to arrive at distinctive features.
- **Model creation:** In this stage, humans are involved in the learning process. [22] has identified three main categories including adding new in-

formation directly, determining and modifying parameters and modifying parameters using parameter learning.

- **Model selection:** Humans can contribute by selecting appropriate algorithms among a set of candidate ML methods according to some criteria. Some algorithms perform better on certain datasets, making it necessary for people to decide which is best to use.
- **ML evaluation & refinement:** Human experts can evaluate and validate outputs of the ML model, which is crucial for the performance of the trained model.

### 2.2.2 Basic principles for designing annotation interfaces

The design of the annotation interface can heavily effect the quality of the annotations, therefore it is important that the basic principles of designing annotation interfaces are discussed. It is important to mention that the guidelines do not apply to all situations, but should be considered based on the task, context and user of the interface. [19] identified four main principles for designing effective annotation interfaces:

- **Cast your problems as binary choices wherever possible:** Presenting problems as binary choices may reduce intra-annotator variability. People appear to be more reliable when asked to rank two items rather than to judge a problem on a continuous scale.
- **Ensure that expected response are diverse to avoid priming:** Order effects and other contextual information might influence the annotation. The most significant priming problem for annotation is repetition. For example, annotators might change their opinion about what is considered negative while annotating the sentiment of social media posts. Ensuring long-enough practice before users start annotating may help them to become familiar with the data so that they have configured there understanding. Another approach is to use a diversity sampling method to make sure each item is as different from the previous one.
- **Use existing interaction conventions:** The interface should make use of basic human-computer interaction conventions, because they have been created by experts and are hard to improve. For example, the interface should let a user know when it pressed a button by providing appropriate feedback.
- **Allow keyboard-driven responses:** It is much slower to use the mouse than the keyboard. In most applications, the Tab key is the designated key to move the cursor to the next field, therefore, it is important to ensure that this is also what happens.

## 2.3 Explainable Artificial Intelligence (XAI)

XAI is an upcoming research area that focuses on the development of AI-systems that can provide explanations for their decisions. AI models such as deep neural networks operate as a "black-box" as it is unclear how it came to a decision. In some cases this is not a problem. For example, when models are very reliable such as in optical character recognition or when mistakes do not carry consequences such as predictions for a movie recommender-system. However, in most cases, a single measurement value is not enough to make real-world decisions. [23] identified several motivations that humans have to make machine learning models interpretable. A few are highlighted below:

- **Scientific understanding:** Humans want to understand how decisions are made so that they can incorporate this knowledge into their lives. For example, business owners might want to know why a product was recommended to certain people so that they can use this intelligence to sell more products.
- **Model debugging and auditing:** An interpretation for an erroneous prediction helps to understand the cause of an error and delivers a direction for how to fix the system.
- **Human-AI cooperation:** Understanding how a machine makes decisions, persuades humans to use the system.
- **High-risk applications:** In some industries it is necessary to provide explanations. For example, a customer has the right to know why he was not granted a loan so that it is clear that it was not due to an unfair process.

### 2.3.1 Taxonomy of XAI methods

Methods for machine learning interpretability can be classified according to various criteria. The information in this section is based on Christoph Molnar's book, "Interpretable machine learning" [23].

**Intrinsic or post-hoc** Intrinsic methods explain machine learning decisions through its inherently simple structure, such as linear regression models or decision trees. For example, to calculate how much a feature contributed to the prediction it is not hard to multiply the feature weight with the desired feature value. However, the interpretability depends on the complexity of the models' internal structure. When a decision tree is very dense, it may still be hard to understand how the decision was constructed. Post-hoc methods are used to explain complex "black-box" machine learning models such as neural networks, but can also be applied to intrinsically explainable models. Another characteristic is that post-hoc methods are applied after the model is trained.

**Model agnostic or model specific** As the name implies, model agnostic methods can be applied to all machine learning models, while model specific methods only to some models. Model agnostic methods are applied after the model is trained, therefore are always post-hoc methods. Model-specific methods use the model internals (i.e., it extracts statistical information from the model itself) to explain how a decision was made. For example, using the internal structure of a linear regression to explain what features contributed to a prediction cannot be applied to another machine learning model. Therefore, model-specific explanation methods are always intrinsic.

**Local or global** Local and global explanations are two different ways of distinguishing between the scope of the interpretation methods. A local explanation only makes one instance or prediction interpretable, while a global explanation method makes an entire model behaviour interpretable.

### 2.3.2 Result of interpretation method

The various interpretation methods can be roughly differentiated according to their results. The information in this section is based on Christoph Molnar's book, "Interpretable machine learning" [23].

**Feature summary statistics** The importance of a feature can be shown as a number, but more complex statistics such as the feature interaction strength are commonly used as well.

**Feature summary visualisation** Feature summary statistics can also be visualised. The advantage is that a large number of data points can be made visible in one central display, making it possible to provide understandable insights of complex spaces. Some explanation methods such as partial dependence plots can only be visualised.

**Model internals** The model internals are similar to their results for inherently interpretable models. For example, the coefficients of a linear regression model, which, in this case is similar to the feature summary statistics.

**Data point** Some explainable methods generate new data points or provide existing ones from the training set to explain a prediction. A Counterfactual explanation is an explanation method that provide the user with new data points. Users may alter features to see how it changes the prediction. For example, to reduce the risk on diabetes, a user may increase the number of exercise hours to how it affects the risk score. Another explanation method that uses data points to explain a prediction are prototype explanations. Imagine a model that predicts the house price for a given apartment. To explain to the user how it came to that prediction, the system provides different apartments with a similar predicted price. This way, the user may pick-up patterns in the

provided examples (e.g., all 5-room apartments are around this price) and may learn the rationale behind the decisions over time.

### 2.3.3 Human-friendly explanations

Explanations are ultimately used by people, so it is important to investigate what makes an explanation "good" for human interpretation. Most explanations are not suitable for laypeople or people with little time. For these people, the explanations often contain too much information or are too technical. To make good explanations, literature from the humanities has been used [5]. The information in this section is based on Christoph Molnar's book, "Interpretable machine learning" [23].

**Contrastive** A contrastive explanation shows the recipient of the prediction the difference between the prediction made and a reference prediction. This way of explaining comes from the way people prefer to receive an explanation, which is best explained using an example. Imagine not being hired for a job you would have liked. When asked why you were not hired, you were given a list of all arguments for and against not hiring you. However, you only wanted to know what you could have done differently to get hired. This list does not show you the difference between the actual result and the desired result, deciphering this costs valuable time. A short statement that manages to name the biggest difference is most appropriate, especially in healthcare where there is little time left to understand explanations of predictions.

**Selective** An explanation of a prediction should contain at most 1 to 3 reasons for explaining an event, even if the event is actually more complex. Often there are multiple reasons for an event, but people are not interested in a complete overview of them. People prefer to see a clear reason for an event, something you also see when watching the news.

**Social** An explanation has to adapt to the environment it is in and the target audience it addresses. If a doctor gives an explanation for starting treatment, the explanation to a fellow doctor will be incomprehensible to the patient.

**Focus on the abnormal** According to Kahnemann & Traversky [24], in an explanation of an event, people usually look for the deviant reason. For explanation design, this means that if there is an anomalous reason it should be shown even if other reasons have more influence on the prediction.

**Consistent with prior believes** Explanations that do not align with people's prior believes are seen as wrong or are being ignored. It is common knowledge that there is a positive causal relationship between smoking and cancer. If an explanation says that smoking has a negative contribution to cancer, no one will trust this model.

## 2.4 XAI-interface design

Determining what constitutes a "good" explanation is often in the hands of the machine learning algorithm developer, someone whose knowledge and background is not representative of end-users' expertise [5]. Developers usually focus on the technical challenges of generating an explanation, but pay little attention to interface design and do not consider the needs of end-users. HCI researchers [6] have developed a framework for designing user-centric interfaces of explanations of ML models (see Figure 1). This framework assumes that the context (who, why, when, where) in which the model is located answers what information should go into the interface and how this information should be presented. This framework will be explained further below.

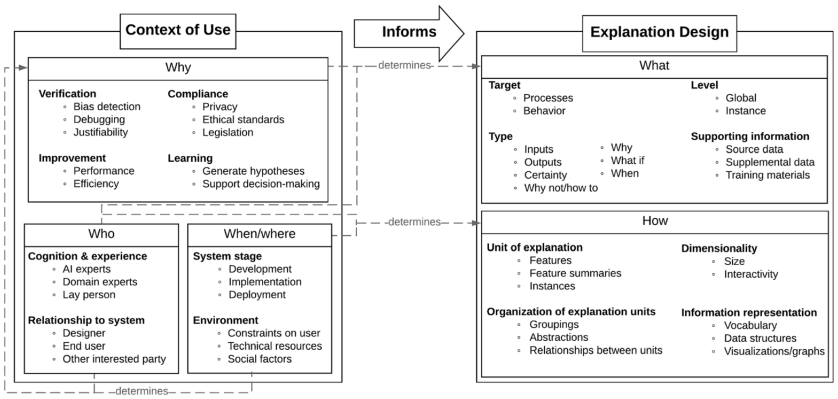



Figure 1: XAI-interface design framework by Barda et al.

### 2.4.1 Context of use

**Who** The framework recognises that users often have multiple roles, so it categorises users across two aspects: 1) user cognition & experience and 2) user relationship to the system. [25] argues that the three groups (developers, domain experts and lay users), differing in terms of user cognition & experience, have different needs in terms of the type of explanation and how it should be presented. According to this study, both domain experts and lay users desire post-hoc explanations, but to the domain expert this should be presented as a visualisation and to the lay-user as a short explanation in text. [26] classifies users (engineer, developer, owner, end-user, data subject, stakeholder) based on the relationship the user has with the AI-system.

**Why** Despite previous studies having made different classifications, according to [6] most goals can be summarised in the following four categories. It is important to note that these goals are not mutually exclusive:

- **verification to the system:** examining how decisions are made by the system to ensure it is operating as expected
- **learning from the system:** extracting knowledge from the system
- **compliance to legislation:** ensuring the system adheres to an established legal, moral or other societal standard
- **improvement of the system:** improving system performance, efficiency, and/or utility

**When/where** The framework distinguishes two aspects on which the environment can be classified: the system stage (development, implementation, and deployment) and the environment stage (constraints on the user, technical resources, and social factors). This is also strongly related to the user's role. A developer in the development phase may desire mathematical insight into the model, while a domain expert may be more in need of how fair predictions are for different demographic groups.

### 2.4.2 Explanation design

**What** An explanation may contain information about a system's internal process or behaviour in general, which in this framework is referred to as the target. Processes explain how the model connects input values to output values, so it exposes the model's internal working and mechanism to the user. Examples include: feature importances, activation patterns in neural networks and decision rules. The system can also explain the general behaviour of the model, which means that it presents the input-output relationships instead of the inner workings. For example, it could explain to users that fever, extreme headache and fatigue are more often associated with a particular condition. Furthermore, an explanation can also differ at the level it is explained. The system may explain the whole system (i.e., global) or one particular prediction of interest (i.e., instance). However, the type and level can usually be determined by the type of an explanation. [27] made a handy overview of the most common types of explanations:

- **"input" explanations:** provide information on the input values being used by a system.
- **"output" explanations:** provide information on specific outcomes, inferences, or predictions.
- **"certainty" explanations:** provide information on why an expected output was not produced based on certain input values.
- **"why" explanations:** provide information on how a system obtained an output value based on certain input values.

- **"why not"/"how to" explanations:** provide information on why an expected output was not produced based on certain input values.
- **"what if" explanations:** provide information on expected changes in output based on certain changes in the input.
- **"when" explanations:** provide information on which circumstances produce a certain output.

The importance of supporting information is emphasised in several studies [7] [25] [6]. The framework underscores source data (i.e., raw data on which the model is built), supplemental data (i.e., data that was not used to train the model on, but is relevant for interpreting the predictions) and training material (i.e., information related to the development of the model).

**How** Within Barda et al.'s framework [6], the presentation of an explanation can be summarized across four key dimensions: 1) the format of the explanation (e.g., raw features, feature summaries, images, or instances); 2) the arrangement of units (e.g., groupings, hierarchical or relational structures, or summary abstractions); 3) the extent of explanation's dimensions, which encompass the overall size of an explanation or interactive exploration options; and 4) the method of information representation, encompassing the vocabulary, data structures, and visualizations employed to convey information. The specific choices made in each of these four primary categories will be influenced by the intended user of the explanation (i.e., who) and the context within which it is being delivered [28].

## 2.5 Collaboration in ICU

Most of the literature on XAI-interface design is focused on single end-users, however care in the ICU is delivered in a highly collaborative and social environment. Therefore, the interface should support this collaborative process. In this sub chapter, a short introduction to collaboration in the intensive care unit will be provided and the importance of supporting this into the development of a novel XAI-interface.

Diagnosing and treating patients in the intensive care unit is complex and challenging that requires many care providers with specialized expertise's. For example, physicians diagnose, treat and manage diseases, nurses are skilled in monitoring and administering medications, microbiologists are specialized in micro-organisms, and infectious disease specialists are specialized in diagnosing and treating patients with infectious diseases. By working closely together, care providers can combine their expertise to share insights, contribute the development of accurate diagnosis and effective treatment plan. According to [14], two types of collaboration exist, both of great importance for the provision of effective and cohesive care.

**Synchronous collaboration** Synchronous collaboration refers to real-time communication among healthcare providers. This often includes immediate information exchange through face-to-face meetings or phone calls. This is very important. For example, when a patient's conditions deteriorates, nurses can contact the designated physician to assess the situation and decide for the most appropriate course of action. This allows for timely adjustments in critical situations, which is crucial as diseases become dangerous within hours.

**Asynchronous collaboration** On the other hand, care providers work together asynchronously when they communicate and share information that do not happen in real-time. Some patients remain for several weeks on the ICU, therefore care providers must provide care in different shifts, thus asynchronous. It is important that care providers are aware of the latest developments so that they can continue the line of work of the previous care providers. This is often the first task of care providers when they start shift. They'll look into the EHR, ask for elaboration with other healthcare providers, during shift hand-offs.

Healthcare providers bring diverse expertise to the table. For example, nurses have close contact with patients, therefore they may detect small changes in the patients' condition faster than physicians. Care providers have domain-specific insights and practical experience. Collaboration allows for the integration of these different perspectives into the XAI-interface, which is needed to make sure all stakeholders can make informed decisions. They have a deep understanding of the clinical context in which the diagnositc prediction model will be used.

## 2.6 Data visualisation

Examining contextual information with huge amounts of data is daunting, therefore visual analytics may come to the rescue. Electronic Health Record (EHR) have been developed to keep patient history and reduce the time spent analysing patient information. This information is very useful, however information is scattered across various sources and information often appears to be out-dated, incomplete and inconsistent [29] [14]. Furthermore, EHRs primarily focus on data storage and retrieval, providing structured representation of patient information. However, they may not effectively convey information in the right format to interpret ML predictions.

In this subchapter, a visualisation overview will be provided based on Tamara Munzner's book "Visualization Analysis & Design" [30].

### 2.6.1 Four stages of validation

In Tamara Munzner's book, she breaks down the visualization design process into four phases: 1) domain situation, 2) task and data abstraction, 3) visual encoding and interaction idiom, and 4) algorithm. The output of each preceding phase feeds into the next. The benefit of this methodology lies in the autonomy of analyzing and validating each phase independently. However, a downside

emerges if any inaccuracies from earlier phases persist, potentially affecting subsequent stages. Therefore, meticulous examination of the output is essential. Despite their apparent sequential arrangement, these phases often necessitate iterative approaches in real-world applications. This subsection will provide a succinct overview of the four phases:

**Domain situation** In the domain situation phase, a picture of the target audience is painted, exploring their areas of interest, questions and problems. The domain usually has its own vocabulary and its own way of solving problems. This can be investigated by conducting interviews with users and observational studies. The outcome is a clear picture of user needs. A common pitfall is making assumptions instead of really engaging with users.

**Task and data abstraction** In the next phase, the domain-specific language is translated into a visualization language. In this way, a designer can determine which processing and coding methods are available and appropriate. For example, by determining whether the data is categorical or ordered, a designer can determine which colors are appropriate in the next phase. Sometimes very different domain-specific situations can be translated into the same abstract tasks.

**Visual encoding and interaction idiom** In this phase, the visual representation of the abstract data block defined in the previous phase is determined. There are two main points to consider when designing. First, it must be determined how data will be presented to the user, for example, what shapes and colors the data points will have. Second, it must be determined how the data can be manipulated, such as whether the information should be able to be sorted or filtered. Although it is often possible to analyze coding and interaction idioms as separate decisions, in some cases these decisions are so intertwined that it is best to consider the result of these choices as a single combined idiom.

**Algorithm** In the last, most nested, phase, an algorithm is developed for the design. The chosen visual representation and interaction method should be handled as efficiently as possible.

### 2.6.2 Marks and channels

After the domain-specific language is abstracted, the abstract information can be converted to a visual encoding. Tamara Munzner provides building blocks for this that can be used to analyze visual encodings. The essence of the design space of visual encodings involves an orthogonal combination of two elements: graphical elements, referred to as markings, and visual channels that control their representation. Even complex visual encodings can be decomposed into components that can be analyzed based on their markings and channel structure. This subsection will briefly discuss the two types of elements:

**Channel types** Channels can be used to manipulate the appearance of markings independent of the dimensionality of the geometric primitive. Like the human perceptual system, Tamara Munzner subdivides channels into two sensory modalities. The identity channels tell us information about what something is or where it is. Magnitude channels, on the other hand, tell us how much of something there is. Examples of visual channels include: shape, the color channel of hue and pattern of motion. Multiple channels can be combined to redundantly encode the same feature. The limitation of this approach is that more channels are "used up," so fewer attributes can be encoded in total, but the advantage is that the attributes shown are very easily perceived.

**Mark types** A marker is a basic graphic element in a visualisation, such as points, lines and areas.

**Channel rankings** Channels can be ranked according to two expressiveness types of ordered and categorical data (see Figure 2). The expressiveness principle stipulates that visual encoding should accurately represent all the information within dataset attributes, without omitting any. This principle is fundamentally demonstrated by displaying ordered data in a manner that aligns with our natural perceptual recognition of order. As can be seen in the picture below, position on a common scale is more effective than color saturation for ordered attributes. This also underscores why barcharts are so widely used.

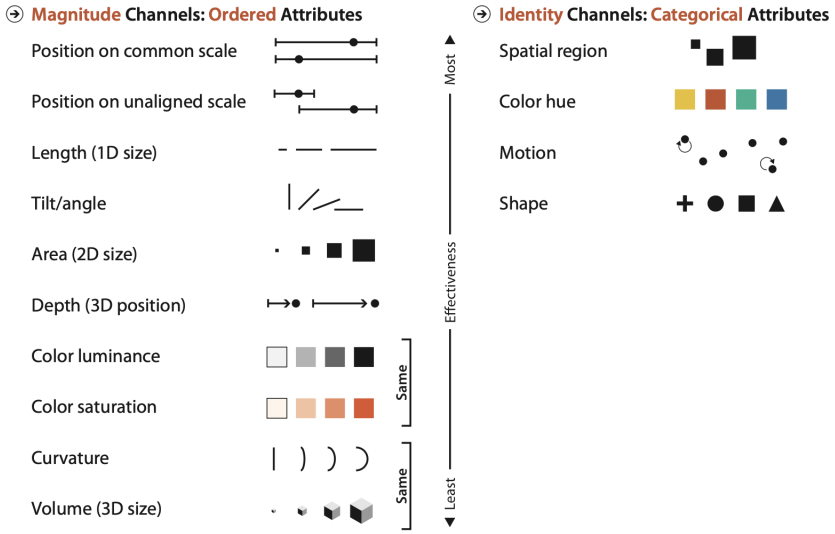



Figure 2: Channel rankings by Tamara Munzner

### 2.6.3 Actions

Tamara Munzner has designed a framework (see Figure 3) with words that describe why people use data visualization to distinguish between different goals. It is divided into three levels of action. At the highest level, visualization is used to analyze information, both consuming and producing it. At the middle level relates to the type of search action (e.g., browsing, exploring). At the lowest level, the goals relate to the type of search (e.g., compare, summarize). Below, the different goals will be described:

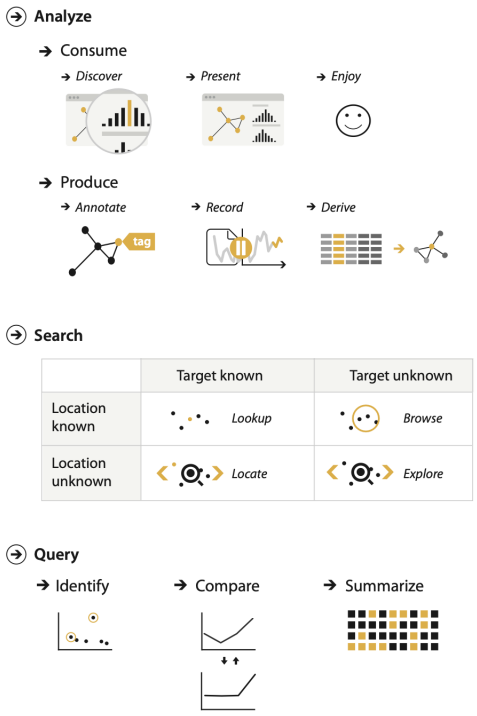



Figure 3: Three levels of actions: analyze, search, and query by Tamara Munzner

#### Consume

- Discover: using visualization to gain new knowledge. This can be completely finding new things (generate hypothesis) or finding out if a conjecture is true or false (verify hypothesis).
- Present: using visualization to communicate something specific to other people. This can take place, for example, within the context of decision-making and planning processes.

- Enjoy: use visualization to entertain users. A visual can elicit curiosity, such as an info-graphic accompanying a blog post.

**Produce**

- Annotate: Adding graphical or textual annotations in a visualization. For example, annotate all points within a text label with a text label.
- Record: Save or capture visualization elements. For example, screenshots, parameter settings or annotation
- Derive: Derive new data points from existing data points by transforming data, for example, by calculating the sum of different variables.

**Search**

- Lookup: Finding information that users know what it is and where to find it.
- Locate: Find information that users know what it is but not where to find it, or in other words, find out where the information is located.
- Browse: Find information that users do not know exactly what it is, but approximately where to find it. For example, the average square meter price in Utrecht on April 20, 2023.
- Explore: Find information that users are not sure what it is and where to find it. For example, outliers in a scatterplot visualisation.

**Query**

- Identify: search that returns the attributes of a single, known target
- Compare: search that compares multiple targets.
- Summarize: provide a comprehensive view of all kinds of different things.

### 2.6.4 Four commonly used EHR-based visualisation types

Rostamzadeh et al.'s study [31] undertakes a review of Visual Analytics (VA) within the context of electronic health records. The study identifies four prevalent categories of visualizations frequently employed in VA systems centered around electronic health records: relation-based, time-based, hierarchy-based, and flow-based visualizations. Each of the categories will be concisely discussed below:

**Relation-based** Relation-based visualizations show the relationships between one or more attributes. A large number of visualization techniques can be used to display relationships: scatter plots, parallel coordinates plots, bubble charts, bar charts, and heatmaps.

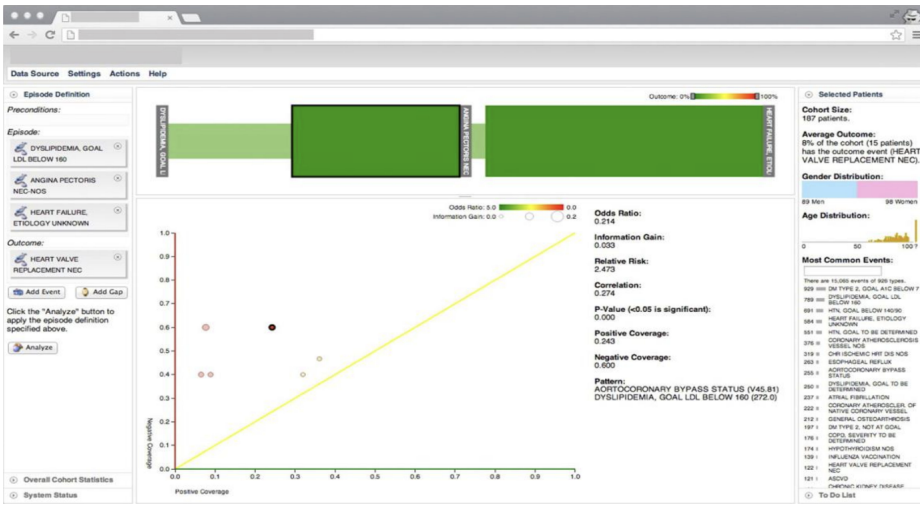



Figure 4: Gotz et al. [32] employ a scatter plot to demonstrate the distribution of the most common patterns with respect to their level of support for various patients.

**Time-based** Visualizations that focus on time display information or events in the order they happened over a period. These visualizations help doctors and experts understand a patient's medical history better. One common way to do this is by using a Timeline. A Timeline shows events in order using icons that can look different in size, shape, or color to show different things about each event.

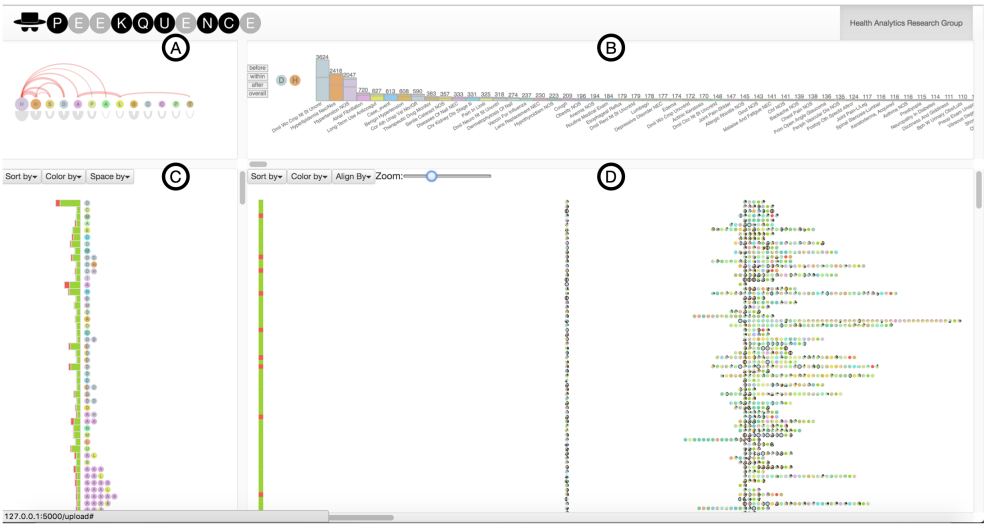



Figure 5: Peekquence [33] displays each patient's event sequence in a timeline.

**Hierarchy-based** These visualizations show how things are put in order and ranked in a system. There are different ways to show this, like using tree diagrams, treemaps, or icicle plots.

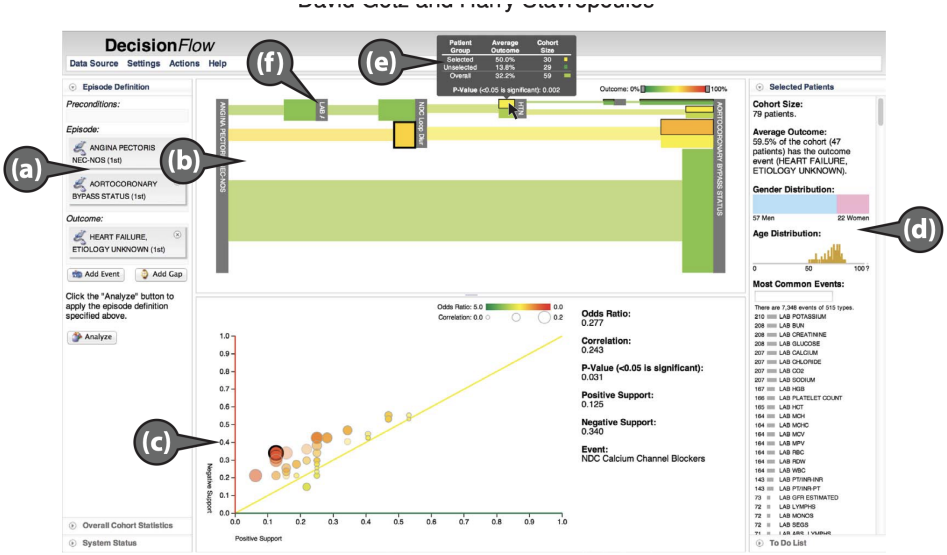



Figure 6: In DecisionFlow [34], they put events that are alike together in a tree. Each spot on the tree represents an event and where it fits in the order of when things happened.

**Flow-based** These visualizations display the movement and amounts of different things compared to each other. Two common ways to do this are using pictures called Sankey diagrams and parallel sets. These are used in systems that look at electronic health records to show how patients move between different types of medical events.


Figure 7: Care Pathway Explorer [35] uses a special kind of diagram called a Sankey diagram to show how clinical events that happen often are linked to each other.

## 2.7 Existing XAI interfaces in healthcare

In the upcoming section, three established XAI interfaces will be examined, with a primary emphasis on their key contributions, dashboard components, and their relevance to the present study, which seeks to explore the optimal contextual information to be presented on the dashboard. To conclude, prevalent design patterns that emerge across these interfaces will be delineated.

### 2.7.1 XAI interface for DCIP Risk Model


Figure 8: XAI interface for DCIP Risk Model [12]

In the study conducted by Bienefeld et al. [12], the primary focus was on understanding user needs to derive insights for the design of an interface tailored for a Delay Cerebral Ischemia Prediction (DCIP) system within the domain of a neuro-ICU. The study provides valuable insights about the aspects of designing an XAI interface by shedding light on information essential for clinicians to interpret ML-predictions.

The core components of the designed interface includes an overview of risk scores (B), incorporating both static and dynamic risk scores. The static scores pertain to factors like patient demographics, while the dynamic scores involve physiological parameters. According to Bienefeld et al. [12], the separation of risk scores allows care providers to discern the relative contribution of each attribute set. However, showing multiple risk scores can also lead to difficulties in interpreting risk scores, which can lead to problems for adopting a model in clinical practice [36]. The dashboard includes visualisations to illustrate both dynamic- (E) and static (C) feature contribution. Static contributions are represented as a bar chart, while dynamic contributions are depicted as heatmaps, providing a representation that matches with the underlying data. Additionally, the dashboard integrates contextual information such as vital parameters (F) and demographic patient information (A). This holistic approach ensures that care providers have access to relevant information, enhancing interpretability.

The outcomes of the study emphasize the clinicians' insistence on the clinical plausibility of model predictions. The presentation of information aligning with clinical knowledge, such as pertinent biomarkers (e.g., C-reactive protein

(CRP)) and physiological parameters, was highlighted as crucial. Clinicians underscored the significance of physical examination findings, underscoring their role in providing essential context. The study suggests that integration with existing systems, such as EHR, is imperative for acquiring patient-specific information from physical examinations.

Furthermore, the study accentuates the clinicians' need for rapid interpretation of model results. In the case of sepsis predictions, the interface's design should facilitate rapid decision-making.

### 2.7.2 XAI Interface for PICU In-Hospital Mortality Risk Model


Figure 9: XAI Interface for PICU In-Hospital Mortality Risk Model [6]

Barda et al. [6] conducted a study with the objective of designing an interpretable interface for a Pediatric Intensive Care Unit (PICU) mortality risk model, with a focus on addressing providers' needs. The mortality risk model is designed to predict in-hospital mortality in the PICU setting, aiming to enhance the acceptability of ML-based systems among clinicians.

An important component of the interface are model explanations, using the SHaply Additive exPlanations (SHAP) algorithm to offer model-agnostic, instance-level explanations. Visualising the feature importances through a tornado plot (B) was found easier to understand than a forceplot, which was consistent with findings from related studies [37]. Additionally, the design emphasized the inclusion of contextual patient information, encompassing raw pre-

dictor values (D), as well as demographic and diagnostic patient details (C & E). Recognizing the critical role of such details in assessing clinical relevance, credibility, and utility of predictions, the study acknowledged their significance in establishing trust in the model predictions.

Despite the initial emphasis on including patient information in the interface, the final iteration opted to exclude this information. This decision was grounded in the anticipation of integrating the model into the EHR system, where contextual patient information is abundant. The study suggests that, while contextual patient information is valuable for verifying predictions and building trust, considerations must be given to the workflow within which the XAI interface is embedded.

The study underscores the significance of user-centered design, with feedback from focus group sessions playing a pivotal role in refining the interface. The primary takeaway from the research is the importance of contextual patient information in verifying predictions. However, the study emphasizes that the integration of the XAI interface into the broader workflow, particularly within the EHR system, should be carefully considered for optimal usability and effectiveness.

### 2.7.3 Sepsis Watch

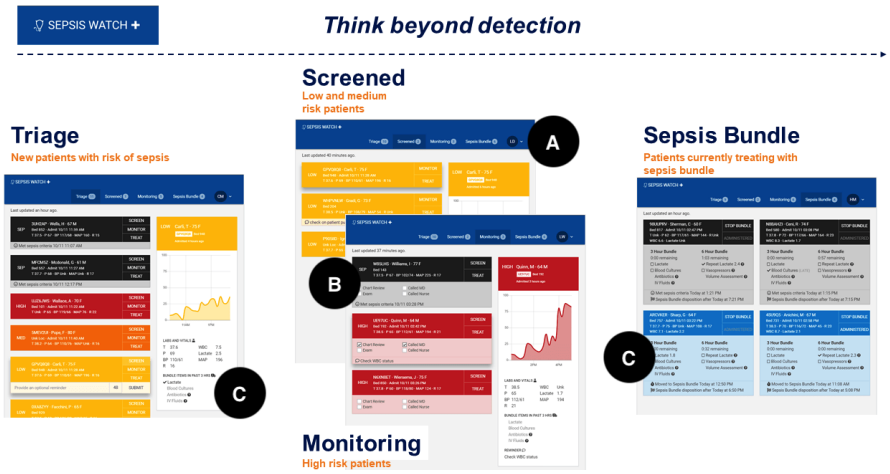



Figure 10: Sepsis Watch Dashboard [9]

This study [9] aims to explore the optimal integration of an ML-based sepsis early warning system, known as Sepsis Watch, into the clinical workflow. Utilizing a deep learning model that analyzes real-time clinical data to predict a patient's likelihood of sepsis, this study is pivotal in discerning the contextual information care providers require and how it should be presented in an XAI

interface for a sepsis prediction model. The research delves into the factors influencing clinicians' acceptance and usability of the ML-based system.

The primary components of Sepsis Watch's main dashboard include an organized overview of patients (B) categorized by sepsis risk using color-coded cards (e.g., black for meeting sepsis criteria, red for high risk, orange for medium risk, and yellow for low risk). Tabs (A) are employed to categorize patients into different groups, such as triaged, screened, and those in the sepsis bundle. Detailed risk analysis and contextual patient information (C) are accessible when a patient is selected, encompassing patient demographics, latest lab and vital signs, and pertinent details about the current hospitalization, including admission time.

The study's key findings emphasize that positive experiences with ML predictions and feedback on model success play a pivotal role in building trust in the system. Moreover, the study underscores the critical importance of understanding clinician perspectives and integrating ML models into the existing clinical workflow. Regarding the display of contextual information, the research highlights the significance of real-time clinical data, demographic information, and details about preexisting health conditions in assessing the patient's baseline health risk. The incorporation of feedback mechanisms, including information on the model's success and specific cases detected, significantly contributes to building trust among clinicians.

### 2.7.4 Common design themes

**Explanations** While the Sepsis Watch interface [9] lacked explicit model explanations, providers expressed a crucial need to comprehend the rationale behind the model's predictions. This difficulty in trusting the model without understanding highlights the critical role of explaining model predictions to foster trust between ML-based tools and healthcare providers. This aspect resonates with current advancements in HCI literature, especially in the field of explainable interfaces.

Studies by Barda et al. [6] and Bienefeld et al. [12] contribute insights into this domain by utilizing model-agnostic instance-level explanations to interpret and validate model predictions. They advocate for the effectiveness of this approach in rendering ML predictions more interpretable. Notably, both studies found that representing feature values through a tornado plot (diverging bar-chart) was deemed easy to interpret. However, they jointly underscore the significance of a user-centric design approach in determining the most effective visualizations that resonate with end-users. The emphasis on user-centred design echoes the HCI principles that tailoring interfaces to meet specific needs and cognitive processes of end-users.

**Contextual patient information** All three studies [9] [6] [12] underscore the importance of incorporating contextual patient information for the effective interpretation of risk predictions. Barda et al. [6] integrated demographics and diagnostic details into their prototypes, while Sandhu et al. [9] expanded

on this by including demographics along with the latest lab and vital signs. Bienefeld et al. [12] presented additional vital sign timelines in their dashboard. Despite their inclusion of contextual information, all studies emphasized the crucial need for seamless integration with the EHR, advocating for parallel use with the model.

In the context of current HCI literature on XAI-interfaces, these finding are consistent with the broader discourse on designing interfaces that use contextual information for interpreting predictions and decision-making. The lack of consensus on what specific contextual information should be shown on the interface, and what information should be deferred to the EHR, resonates with ongoing discussions in HCI about optimizing information presentation. This challenge highlights the importance of HCI principles in designing XAI-interfaces that align with the current workflows.

**Supporting model information** Barda et al.'s study [6] found that presenting additional information, such as tables of raw feature values and time-series plots, proved beneficial for healthcare providers in interpreting predictions and explanations. However, showing raw feature values may become problematic for models with multiple features. As emphasized in [23], the importance of simplification and abstraction to enhance interpretability of complex models.

**Risk representation** All three studies emphasize the crucial role of risk representation in facilitating effective interpretation; however, they present varying results. Bienefeld et al. [12] adopted a dual approach, displaying risk scores both numerically (as probabilities) and visually (as a risk analysis with highlighted high-risk areas). This dual representation strategy enhances information comprehensibility, catering to diverse cognitive styles among users, as discussed in the current state of art in HCI literature.

In a similar vein, Barda et al. [6] explored clinicians' preferences for risk information represented as odds versus probabilities. The findings revealed a unanimous preference for probabilities, attributed to the reduction in information processing effort. This aligns with HCI literature, which often advocates for representations that minimize cognitive load and enhance user understanding.

Sandhu et al. [9], on the other hand, discovered that representing a risk score through color-coded categories (low, medium, high) was more intuitive and easier to grasp than depicting it as a continuous risk scalar. This finding resonates with the ongoing discourse in HCI literature, emphasizing the importance of intuitive visualizations that facilitate quick and accurate comprehension of information.

According to these studies, the choice between numerical, visual, or categorical representations should consider the varied cognitive styles of healthcare providers, ensuring that risk information is presented in a manner that optimizes understanding and decision-making.

**Interactivity** The interfaces in these studies incorporated interactive components, such as shared views and on-demand information retrieval, aiming to alleviate cognitive load. This design choice aligns with principles discussed in the current state of art in visualisation and HCI literature. As exemplified by Munzner [30], emphasizes the importance of interactive elements to enhance user engagement and comprehension.

One prevalent design strategy employed in these interfaces is progressive disclosure, a technique that gradually presents information and unveils additional details upon user request. This approach aligns with HCI principles that advocate for managing cognitive load by initially providing a concise and simplified view of the data. In the context of XAI interfaces, especially in the intensive care unit where rapid decision-making is imperative, progressive disclosure becomes particularly crucial. Clear and concise information presentation emerges as a key factor in expediting decision-making processes and reducing cognitive workload, an aspect underscored by the broader HCI literature focusing on user-centered design and effective information visualization.

# 3 Sepsis prediction model

The model that is used for this study is an early warning LOS prediction algorithm. It utilizes low-frequency heart rate and oxygen saturation data obtained from the neonatal intensive care unit (NICU). This chapter discusses the main model limitations by placing these in the context of XAI interface design. In addition, it provides an overview of all the predictors used in this model. More information about the model can be read in their published article [38].

## 3.1 Model lmitations relevant for XAI-interface design

**Limited Variables** Some predictors appeared to be unreliable and were excluded from the model. Variables such as patient temperature, commonly used to identify fever, a symptom of sepsis, were omitted due to influences like incubator temperature and measurement accuracy. Skin color, another commonly used factor, was not included as the child is often covered in the incubator. Additionally, infection parameters like CRP were considered too delayed for an early warning system. The model utilized heart rate and oxygen saturation as predictors. Transparency regarding the model's capabilities, as highlighted in XAI literature [39], is essential for building trust and making informed decisions. Therefore, the interface should explicitly state the predictors used by the model.

**Use of Low-Frequency Data** The prediction algorithm uses low-frequency data to make predictions, impacting model sensitivity. It may not capture subtle and rapid changes in physiological parameters, failing to detect short-term trends and variations. This limitations can results in a reduced sensitivity to early signs of medical conditions, potentially leading to missed diagnosis. A high number of false negatives could become a problem, especially when care providers over rely on the model.

**High Number of False Positives** Despite the model's moderate performance (AUC of 0.73 upon clinical suspicion), a notable issue is the relatively high number of false positives, indicating that high-risk predictions may have only a minimal chance of actual sepsis. The abundance of false positives emphasizes the need for contextual information, enabling care providers to assess the clinical relevance of alerts and make informed decisions. Additionally, selecting the appropriate alarm threshold is crucial for model adoption; an excessive number of false positives can erode trust in the model. To address these concerns, incorporating a feedback mechanism within the interface, allowing users to provide feedback on false positives and other observations, can contribute to continuous model improvement and refinement.

**Cross-sectional dataset limitations** The sepsis prediction model underwent training using a cross-sectional dataset characterized by 4-hour interval

time-aggregated features (mean, minimum, variance). Consequently, the model may encounter challenges in fully capturing the temporal dynamics and correlations present in repeated measurements, which are inherently a part of longitudinal data. This limitation has the potential to influence the prevalence of positive LOS cases, deviating from what would be observed in a purely cross-sectional analysis.

Furthermore, the absence of intricate temporal dynamics and correlations in repeated measurements introduces difficulty in calibrating the model to align its certainty with the actual likelihood of a patient having sepsis. Consequently, interpreting risk scores derived from the model becomes a nuanced task, as these scores cannot be directly translated into the true probability of a patient having sepsis. In essence, the model's training on time-aggregated, cross-sectional data restricts its ability to capture the evolving nature of sepsis risk over time, limiting the direct interpretability of risk scores as accurate representations of the actual chance of sepsis occurrence.

## 3.2 Main Features Used in the Model:

As explained, [38] states that a range of monitoring data was initially evaluated, but heart rate and oxygen saturation measurements were selected for the algorithm. The main features used in the logistic regression model include:

| Name | Description |
|---|---|
| HF mean | Mean heart rate |
| HF variance | Heart rate variance |
| SpO2 variance | Oxygen saturation variance |
| SpO2 min | Minimum oxygen saturaion |
| Bradycardia | Number of bradycardias |
| Tachycardia | Number of tachycardias |
| SpO2 drops | Number of oxygen saturation drops |

Table 1: Selected features for the sepsis prediction algorithm [38]

# 4 Design study framework

Sedlmair et al. [40] introduced a framework comprising nine phases organized into three categories for carrying out a design study (see Figure 11). The "precondition" phase outlines the necessary preparatory steps before commencing a design study. The "core" phase details the sequence of actions involved in conducting the study, while the "analysis" phase involves presenting outcomes and reflecting on the design study. Although the framework is presented as a sequence, it doesn't mean each previous step 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. The following sections will discuss an outline of the framework, with a detailed description of each phase available in Sedlmair et al.'s study [40]. Subsequent chapters will delve into the discussion of each phase concerning the design of an explainable interface for the sepsis prediction model. The deploy phase is omitted due to time constraints, as the interface will not be deployed.

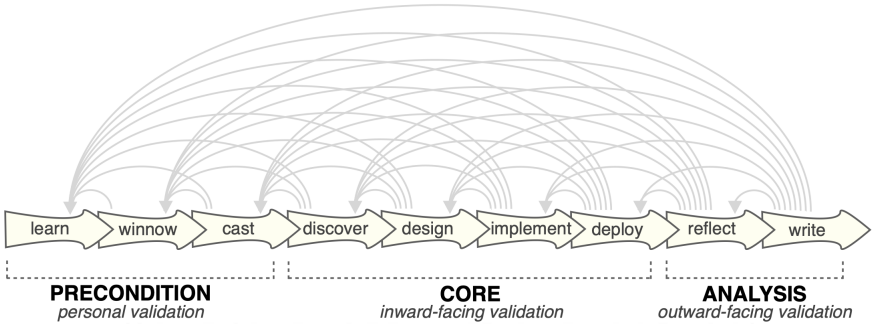



Figure 11: Sedlmair et al. [40] nine-stage design study methodology framework organized into three top-level categories

## 4.1 Precondition phase

The initial steps of learn, winnow, and cast revolve around getting the visualization researcher ready for the task and identifying and selecting collaborative partnerships with domain experts.

### 4.1.1 Learn: Visualisation literature

This 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.

### 4.1.2 Winnow: Select promising collaborators

The winnow phase focuses on selecting promising collaborations. This involves initially 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, for example, the time available between two parties.

### 4.1.3 Cast: Identify collaborator roles

This phase defines the roles within the project, with the most important roles being the front-line analyst (domain expert end user) and gatekeeper (project approver), along with additional roles such as connectors and translators. However, these roles are not set in stone. For this project, other roles (e.g., students, researchers) are more suitable.

## 4.2 Core phase

The core of a design study contains four stages: discover, design, implement, and deploy.

### 4.2.1 Discover: Problem characterization & abstraction

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.

### 4.2.2 Design: Data abstraction, visual encoding & interaction

In the design phase, the data abstractions, visual encodings and interaction mechanism will be generated and validated. After requirements are identified in the previous phase, multiple solutions will be generated, leading to one solution. A common pitfall is to choose a solution too early. This can be avoided by generating a wide selection of solutions. The research can incrementally refine the wide range of solutions by using design principles and guidelines. The proposed solutions in the proposal space should be presented to domain experts for discussion, such as in the form of paper mock-ups, data sketches or low-level prototypes. The goal of the design cycle is to satisfy rather than optimize: while there is usually no best solution, there are many good and okay solutions.

### 4.2.3 Implement: Prototypes, Tool & Usability

Once a design solution has been identified in the preceding phase, this stage will involve creating a prototype. Subsequently, this prototype will undergo testing

with the actual users.

#### 4.2.4 Deploy: Release & Gather Feedback

In the final core phase, we put the tool to real-world use and gather feedback. The main goal is to see if experts find the tool helpful.

### 4.3 Analysis phase

#### 4.3.1 Reflect: Confirm, Refine, Reject, Propose Guidelines

Reflection is a crucial part of a design study. This helps build knowledge and lets other researchers learn from the work. It's especially useful for improving design guidelines. When new things are discovered, new guidelines can be confirmed, refined, extended, or even proposed.

#### 4.3.2 Write: Design Study Paper

Writing is about communicating findings with the research community. It is important not to rush but to take time to reconsider ideas and explain them clearly.

# 5 Winnow & Cast

This chapter defines the different roles involved in this study. A role may be separate from a person, as a person can also have multiple roles.

**Researchers** The researchers involved in this project are myself and my supervisors. My supervisors work at Utrecht University, department of human-computer interaction, and my supervisors from UMC Utrecht work in the department of neonatology and digital health.

**Domain Experts** The domain experts involved in this project are the following healthcare providers: neonatologists, NICU nurses, NICU physician assistants and fellow neonatologists. They can be contacted for an interview with permission from my UMC Utrecht supervisor.

**Students** Master students from the Human-Computer Interaction course also participate in the study. They are tasked with providing interface feedback in the design phase of the study.

# 6 Discover

In this chapter, we will explore the domain and existing practices, addressing the challenges and requirements of the target group. The identification of these aspects was accomplished through interviews with domain experts and a half-day observation at the NICU ward. The insights gained from these activities were then transformed into general tasks, which were subsequently validated with stakeholders.

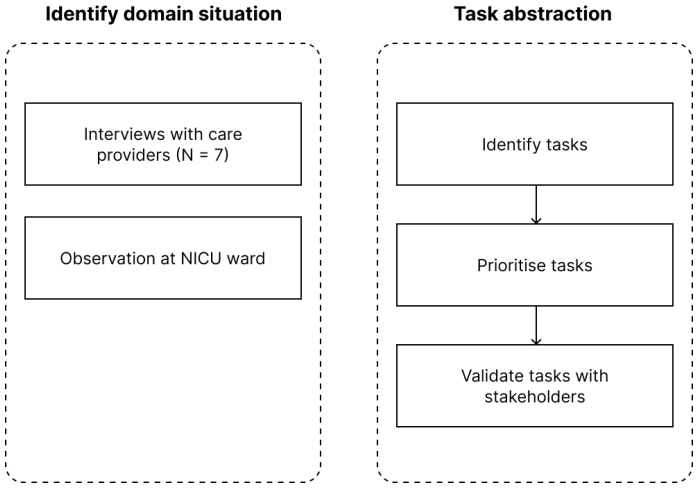



Figure 12: Overview of the discover phase: transforming findings from interviews with domain experts and an observational study into a set of potential requirements.

## 6.1 User Interviews

The first step to identify the needs of end users is by engaging with them. Semi-structured interviews were conducted with four different healthcare providers (M = 1, F = 3). Each interview was conducted independently and lasted approximately 30 minutes. The interview began with a brief introduction of the purpose of the study. It was made clear to the participant that the interview would be recorded, the information would be kept confidential and could be stopped at any time without having to provide a valid reason. Participants were then asked to sign a consent form (see Appendix E). 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 outline of the interview can be found in the Appendix A. A second round of interviews was conducted to identify the exchange of information between caregivers. Four caregivers participated in these, including 3 nurses and 1 fellow neonatologist (M = 1, F = 3). During the interview, sharing information with and retrieving from other care providers

was discussed. The outline of the interview can be found in the Appendix B.

## 6.2 Observational study

In addition to conducting user interviews, half a day was dedicated to observing care providers on the NICU ward. Throughout the observation, valuable insights were gathered pertaining to decision-making, collaboration, and information exchange. The day commenced by following a neonatologist during their morning routine, which included patient visits and offering guidance to fellow care providers. Subsequently, nurses were observed while delivering care, providing an opportunity to pose questions for clarification of their actions. The neonatologist demonstrated the utilization of Metavision for patient administration. Towards the end, participation in a multi-disciplinary session allowed witnessing decisions being made regarding the treatment of challenging cases. Throughout the observation, detailed notes were taken and later expanded upon for comprehensive documentation.

## 6.3 Set of Requirements

The interviews with care providers and related work have yielded the following four themes: explanation, patient information, model feedback & collaboration, and alarm notification.

Care providers indicated reluctance to make decisions based on a prediction alone. They said that they would like to know what the prediction is based on (requirement: 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 (requirements: 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. On the other hand, care providers found these metrics too challenging to understand. Since the large majority of care providers did not want to see this on the dashboard, it was decided not to include it. In addition, as explained in the model limitations in the cross-sectional dataset limitations 3.1 section, calculated likelihoods are difficult due to repeated measurements.

Care providers indicated the need for contextual information to interpret the model's risk score, enabling them to determine the appropriate course of action. For instance, *"in a severe illness, you might need to start treatment at 20%, while in a mild care, you can wait until 80%".* Another care provider said: *"its orange (indicating a moderate risk score), but it's an extremely premature case, so you need to start now".* They indicated that contextual information is key to facilitate informed decision-making. Furthermore, they desired contextual information to verify and validate the model based on their own insights and knowledge. They wanted this information to cross-verify predictions, as one care provider mentioned, *"that you can compare it with the patient. Yes, the*

*child has a lot of bradycardia. Let's start treatment".* They emphasized the importance of having contextual information to enhance their confidence in the model's 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 rapports, resulting in instances where crucial details are overlooked, and takes longer to attend the patient. As one provider noted, *"It is sometimes a lot of information, which causes you to have an overwhelming amount of data to process. This leads to delays before we can actually get to the patient.".*

The added value appears to lie in having all relevant information in one central location rather than solely in the algorithm itself. However, centralizing all information relevant to interpreting sepsis risk scores would enable providers to act quickly and efficiently without spending time searching for scattered information across multiple sources. This enhances care providers ability to interpret risk scores but also improves the overall decision-making process.

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, which will be explained below.

Actual interventions like antibiotics 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 (requirement: 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 (requirement: 2.2). Furthermore, laboratory results like white blood cell count and CRP values serve as indicators for infections in the body (requirement: 2.3). Vital functions provide indicators of the vital organs performance, with heart rate and oxygen saturation being the most critical ones (requirement: 2.4). They are also the input features for the prediction model. Healthcare providers expressed a desire to know the patient's weight and age (requirement: 2.6). Knowing that the child is extremely premature, treatment may be initiated earlier and a moderate risk score could be highly alarming.

The literature shows that providing feedback on a prediction has a positive effect on usability (requirement: 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 (requirement: 3.2).

Care providers wanted an alarm to be forced on them so that no alarms

would be missed (requirement: 4.1).

### 6.3.1 Participants

| ID | Job function |
|---|---|
| P1 | Neonatologist |
| P2 | Neonatologist |
| P3 | Neonatologist Fellow |
| P4 | Infectiologist |
| P5 | Nurse |
| P6 | Nurse |
| P7 | Nurse |
| P8 | Literature [6] [7] [12] [41] |

### 6.3.2 Priorities

| | |
|---|---|
| Must have | Critical requirements |
| Should have | Important, but not necessary for the prototype to be validated. |
| Could have | Desirable, but not necessary. Can be included if time and resources permit. |
| Won't have | Least-critical |

### 6.3.3 Functional requirements

**Theme 'Explanation'**

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

**Theme '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 |

**Theme 'Model feedback & 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 |

**Theme 'Alarm notification'**

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

**Non-functional requirements**

| ID | Requirement | Participant | Priority |
|---|---|---|---|
| 4.1 | The system should minimize cognitive load. | L | |
| 4.2 | The system should group similar objects | L | |
| 4.3 | The system should use simple terminology | L | |
| 4.4 | The system should be providing appropriate feedback | L | |
| 4.5 | Visual information should be supported with textual information | L | |
| 4.6 | The system should minimize the number of actions to complete a goal | L | |

# 7 Design

After identifying the needs of healthcare providers and transforming them into a set of requirements, several low-fidelity designs for the interface were created. 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, together with the XAI and visualisation literature, a set of relevant design solutions were created (see Figure 13). These designed components were tested with end users and usability experts. This chapter will layout the design considerations for the low-fidelity prototypes and detail the results of the evaluation.

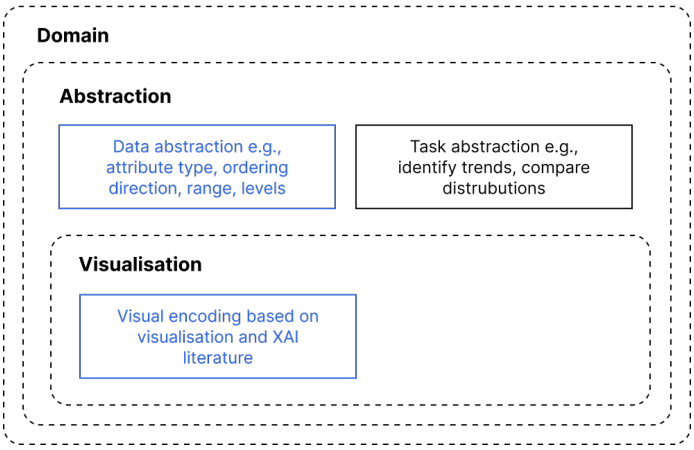



Figure 13: The analysis model by Tamara Munzner [30], in which the current steps of the process are highlighted in blue.

## 7.1 Low-fidelity Prototype

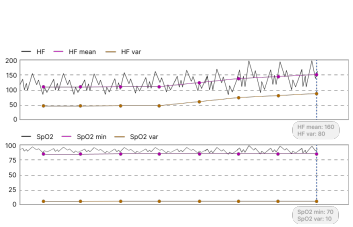
Low-fidelity prototypes are simple, rough and often quick designs of a product or interface. The focus is on the concept and not the detailed graphics. This ensures that concepts can be tested quickly to validate hypotheses and gain valuable insights. A low-fidelity prototype can be created in a variety of ways. This study will use Figma [42], a Web application for designing interfaces. They offer a large library of drag and drop components, allowing layouts to be created quickly and easily. In addition, the designs can be made easily clickable. This allows users to use the interface in a "real-world" way, leading them to different insights.

The low-fidelity prototypes were designed based on Tamara Munzner's book 'Visualization Analysis and Design' [30] and relevant XAI literature. To determine an appropriate presentation form for the set of requirements, the following components were identified: vital sign chart, feature importance chart, risk analysis, risk score, table with patient information, patient overview section, and a

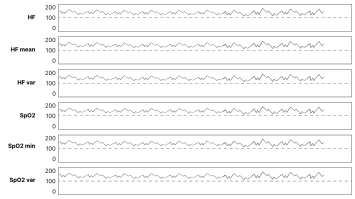
feedback form. For some of these components various relevant solutions were found, which will be further explained below.

### 7.1.1 Vital signs

Two variants for representing the vital signs were designed: a line chart featuring multiple attributes (Figure 14a) and a multi-plot line chart (Figure 14b), one for each attribute. The attributes included all features that were used by the model, excluding countable predictors: heart frequency, saturation, mean heart frequency, heart frequency variance, saturation variance, and minimum saturation. Tamara Munzner [30] 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 (time). However, when confronted with numerous attributes, the chart can become challenging to interpret, leading to the creation of a multi-plot line chart. Despite its benefits, this approach introduces the challenge of comparing attributes effectively.




(a) Variant 1: Linechart with vital signs




(b) Variant 2: A multi-plot linechart with vital signs

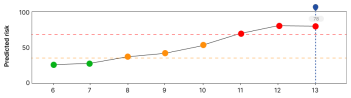
Figure 14: Vital sign sketches

### 7.1.2 Feature importance chart

In designing the feature importance chart, three relevant design solutions were found. The most prevalent visualization was the tornado plot, where factors contributing to an increased risk score are positioned on the right side, while those diminishing the score are located on the left side. This design aligns with the latest XAI and visualisation literature, highlighting the ease of interpretation [43] and suitability for comparing diverse attributes [30]. This makes the tornado plot a relevant design choice for visualising feature importances (Figure: 15b).

Notably, tornado plots were selected over alternative visualizations such as force plots, which were find hard to interpret for individuals lacking AI expertise according to findings by Haas et al. [37] and Barda et al. [6]. 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. Despite the potential utility of scatter plots in revealing how each feature behaves globally, their interpretational challenges, particularly among lay-users [43], rendered them less suitable for quick comprehension.

Interestingly, XAI research [23] indicated a divergence in preference among healthcare professionals, with nurses favoring textual explanations and doctors finding visualisations more interpretable. For this reason, textual explanations (Figure 15c) were also included as a relevant design choice. A hybrid solution was also developed, presenting a combination of textual and visual explanations in a tabular format. Although this approach is visually less telling, it facilitates rapid understanding, especially when dealing with a limited number of features (Figure 15d). Tables offer the added benefit of allowing items to be sorted, enhancing the user's ability to focus on specific aspects of interest. This strategy, rooted in both XAI principles and user feedback, underscores the importance of tailoring explainable interfaces to diverse user preferences and expertise levels.




(a) Prediction trend chart


(b) Variant 1: feature value contribution tornado plot

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) Variant 2: feature value contribution full text

| 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) Variant 3: feature value contribution table

Figure 15: Model information sketches

### 7.1.3 Risk scores

Turning our attention to the risk score representation, two variants were created. One with categorical labels (Figure 16a) and another with numerical values (Figure 16b). Despite the prevalent use of numerical values in existing XAI interfaces [12] [6], categorical representations were found easier to interpret [9] [10]. These findings were consistent with results obtained from our own user interviews, which revealed that care providers had a preference for a categorical approach. They expressed a tendency to interpret parameters by categorizing

them into three groups: "low," "medium," and "high," akin to a traffic light system. Participants emphasized the challenge of deriving practical insights from numerical scores, questioning the meaningful distinction between scores like 80 and 85. Because risk scores are often expressed in a number, both variants are designed, so that they can be tested in the next phase.


(a) Variant 1: risk score with categories (b) Variant 2: risk score with numbers

Figure 16: Risk score sketches

### 7.1.4 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 [44], groups of similar elements were created to enhance interpretability. 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.

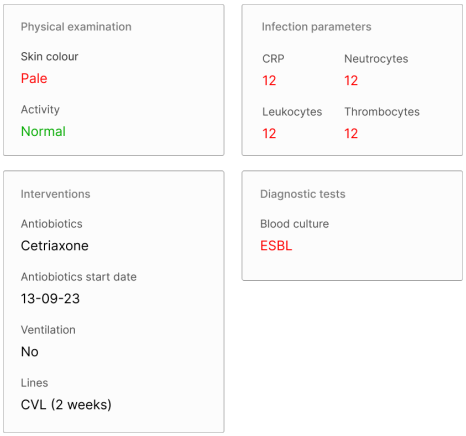
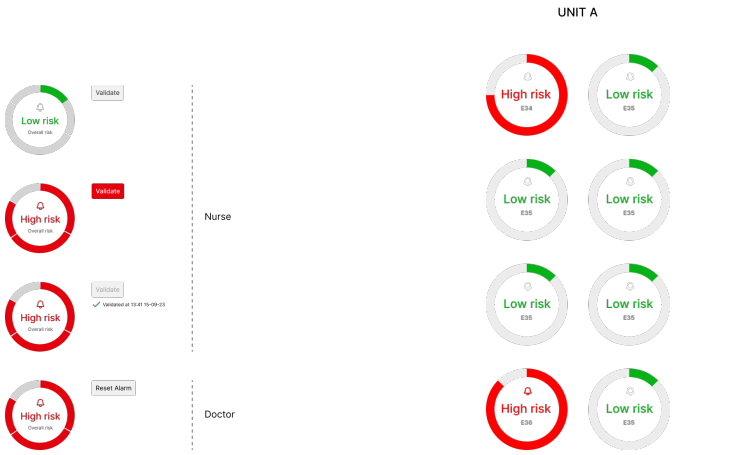



Figure 17: Enter Caption

### 7.1.5 Patient overview

The risk scores were used to create an overview of all patients, with a title specifying the relevant unit (Figure 18b). During the interviews with care providers, as detailed in the subsequent section, care providers indicated a need to validate and mute alarms (Figure 18a). Nurses, for instance, should 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.




(a) Muting and validating an alarm

(b) Patient overview per unit

Figure 18: System component sketches

### 7.1.6 Feedback

According to the human-in-the-loop literature [19], feedback options should be shown as binary choices wherever possible. They 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.

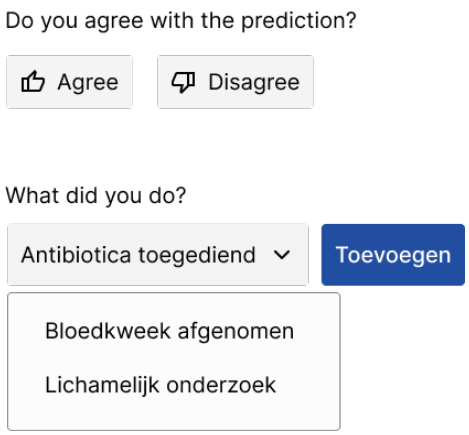



Figure 19: Low-fidelity feedback-form

## 7.2 Interviews with Domain- and usability experts

Domain experts were asked to examine a range of low-fidelity prototypes, aiming to confirm the alignment of the designs with their requirements. They 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. The full interview outline can be found in the appendix C. Furthermore, the designs were critically evaluated by usability experts with the goal of improving the usability of the designs.

### 7.2.1 Participants

The low-fidelity prototypes were evaluated by 6 domain experts, 1 data scientists and 2 usability experts (N = 9, M = 6, F = 3). From the group of domain experts, the following care providers took part: 1 fellow-neonatologist, 1 neonatologist, 2 nurses and 2 physician assistants. All care providers possessed considerable expertise in diagnosing and treating neonatal sepsis. They exhibited diversity in terms of age and practical experience. The usability experts were both second-year MSc. HCI students, providing both practical and academic insights. They both had significant expertise in UX design and were involved in at least one design project within the last two years.

### 7.2.2 Materials

Participants participated in the interviews either in person at the designated location or, if in-person attendance was not feasible, the session was held online via Teams. For participants that attended the session in-person, a private room at the WKZ was made available. For these sessions, the low-fidelity designs were printed out and additional pen and paper was taken to the session. Participants that attended the session online, a Figma file that contained the low-fidelity

prototype was shared. The answers to the interview questions were recorded on a mobile device.

### 7.2.3 Procedure

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 (e.g., participants can stop at any time) and were asked whether they agree to an audio recording. The full informed consent can be found in the appendix E. Domain experts were then asked to explain what the visualizations on the dashboard represent. Next, domain experts were asked which visualizations they found most understandable and useful (i.e., visual encoding and interaction mechanisms) for estimating the risk of sepsis. Usability experts were asked to share one positive and one negative comment for each of the components. 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. The full interview outline can be found in the appendix C. During the interview, participants were promoted to use pen and paper, for example, to draw alternative visualisations. Afterwards, the researcher summarised the main insights, which participants could then confirm.

### 7.2.4 Analysis

The audio recordings were listened back and then deleted. All findings for each participant were noted. Then all duplicate findings were extracted and the remaining findings were sorted from most important to least important for estimating the risk of sepsis. Feedback not relevant to the dashboard was removed from the list and saved for future work. In addition, 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.

### 7.2.5 Results

In this section, feedback from domain and usability experts will be discussed for each of the designed low-fidelity components. The main findings were listed in the following Table 2.

| Component | Finding | Design recommendation |
|---|---|---|
| Vital sign chart | Desire to compare attributes | Show all relevant attributes in the same chart |
| | Used chart to detect outliers | *Highlight outliers with visual cues* |
| | Used chart to identify trends and baselines | *Allow zooming* |
| | Redundant attributes were distracting | Remove redundant attributes |
| | | |
| Patient information | Desire to identify trends and baselines | Display how measurements changed over time |
| | Needed timestamps to assess relevance | Include timestamps & remove outdated measurements |
| | Missed patient information to interpret and validate patient risk | Include missing patient information (e.g., temperature) |
| | Color-coded values were interpreted as feature contributions | Use alternative visual cues than color (e.g., bold text) |
| | | |
| Feature importance chart | Providers had opposing design preferences | *Support multiple visualisations to suit different users* |
| | Desire to identify trends and baselines | Make chart available for each prediction |
| | Varied opinions regarding utility | Make chart available on demand |
| | | |
| Risk trend & scores | Numerical values were interpreted as sepsis likelihoods | Express risk as a category |
| | Desire to identify and baselines | Allow zooming |
| | | |
| Patient overview | Caregivers linked actions to categories, whereas the purpose of the model is to visit the patient. | Only visually distinguish between alarm/no alarm |
| | Desired alarm validation and muting | Allow alarm validation and muting |
| | Providers were only interested in patients from their unit | Show patients within a unit |

Table 2: Overview of the obtained findings from both domain and usability experts. Findings marked in italic were not implemented in the high-fidelity prototype.

**Dashboard composition** Domain and usability experts agreed on the layout of the dashboard for the most part. 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 function charts. Care providers thought the vital signs chart was more important than the risk analysis chart and should therefore be placed lower than the vital signs chart.

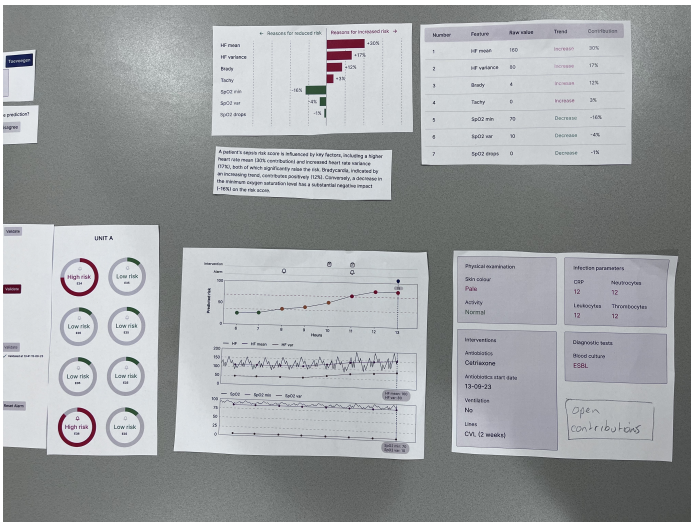



Figure 20: Dashboard lay-out composed by one the participants

**Vital signs** Care providers showed a preference for the line chart featuring multiple attributes in a single plot, as opposed to the multi-plot line chart. They found it more user-friendly for comparing attributes within one consolidated view. Additionally, they suggested the removal of specific attributes, such as 'heart frequency variance,' 'oxygen saturation variance,' and 'minimal oxygen saturation.' According to their feedback, these attributes could be derived from the heart rate and oxygen saturation signs, leading to unnecessary clutter in the plots.

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. Although this feature was taken into account in the high-fidelity design, there was not enough time to implement it in the high-fidelity prototype.

**Patient information** Their feedback extended to the latest patient information parameters, with a specific request for the inclusion of the time elapsed since the measurements were taken. They highlighted the importance of this temporal context, noting that outdated measures might no longer be relevant.

Interestingly, there was 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 model's 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 fellow's and physician assistants) showed a preference for visual representations, particularly the tornado plot format. 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 that understanding the model's functioning was more important for doctors.

**Risk scores** Both domain and usability experts thought the risk categories were easier to interpret. Healthcare providers incorrectly interpreted the model's output as directly related to the likelihood of sepsis. Even in cases of high probability predictions, the actual chance of infection is considerably lower. A score of 75 does not imply a 75% chance of sepsis; in reality, this probability is much lower. As explained in the section outlining the limitations of the model, achieving calibration of the model's output to accurately reflect the true probability of sepsis poses a challenge. This difficulty arises from the presence of repeated measurements, particularly considering that the model is trained on cross-sectional data.

Moreover, displaying the exact probability of sepsis to healthcare providers may not be desirable, as a low probability might not be taken seriously. The difference between, for example, a 3% chance and a 0.01% chance is significant,

but healthcare providers might not perceive it in that way. To make the results more accessible, an approach categorizing outcomes into understandable levels, such as low, medium, and high-risk areas, proves [10] to be an effective solution. 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 area?"*. The model is not distinctive enough to directly link actions to specific risk areas. This means the model cannot indicate when to alert a doctor when the risk area is classified as medium. While the model generates valuable alerts, each alarm signal remains an area with nuances.

Nurses sought practical guidance: *"What steps should I take when an high-risk alarm is triggered?"* In contrast, doctors were more inclined to comprehend the underlying processes. 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.

**Patient overview** Some domain experts showed a positive response to the patient overview component. They thought the information was clear and well structured, indicating helpful for informed decision-making. Usability experts recommended expanding the patient information displayed 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 patient overview intentionally does not differentiate between low, medium, or high-risk areas. The system simply uses two colors: red for an alarm signal and a neutral color for no alarm signal. 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. This feature is crucial as it enables healthcare providers to understand the trend value and identify the patient's context. A transition from low to high risk may potentially be more alarming than a shift from medium to high. The boundaries for the categories are carefully determined by the model developers and depend on the number of alarms they want to be triggered. A higher threshold will result in higher precision, leading to fewer generated alarms.

## 7.3 High-fidelity prototype

A high-fidelity prototype is an advanced, detailed and interactive model of a product or system that accurately mimics the final look, feel and functionality of the final product. This type of prototype is often developed using technologies and tools similar to those of the final product, providing a realistic user experience.

After the data was collected from the interviews, a high-fidelity web application prototype was developed using JavaScript, HTML, and CSS. The interactive charts were developed using D3.js. As we were not allowed to use EHR of even pseudonymous data, mock EHR data was used by anonymizing and scrambling.

In this section, the high-fidelity prototype 21 will be discussed, explaining the dashboards main components.


Figure 21: 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.


Figure 22: Design of the high-fidelity feature value contribution chart

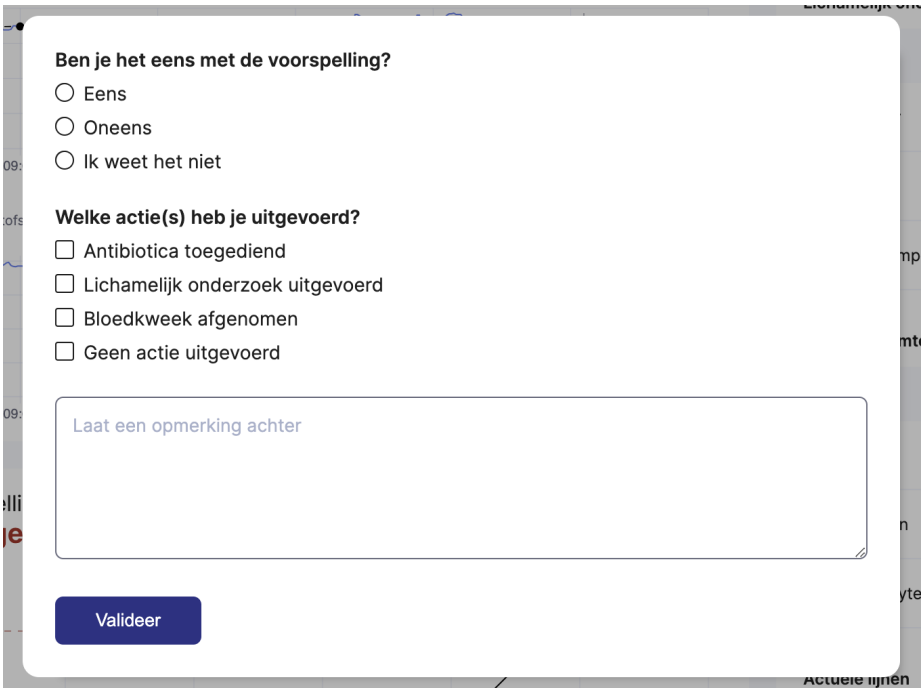



Figure 23: Design of the high-fidelity validation form

### 7.3.1 Patient overview (C1)

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 patient's validation status. For patients without generated 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.

### 7.3.2 Alarm notification (C2)

When an alarm is triggered, the system records the time of the alarm generation. 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. Once validated, the system displays the time at which the patient was validated. The actions performed

can be viewed by other care providers by hovering over the document symbol in the trend prediction chart. It's essential to note that only a physician has the authority to deactivate an alarm. This precautionary measure ensures that decision-making regarding alarm signals occurs responsibly. 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.

### 7.3.3 Vital signs (C3)

Users have the capability 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's location. To streamline the charts and enhance clarity, metrics such as 'heart frequency variance,' 'oxygen saturation variance,' and 'oxygen saturation minimum' were eliminated, 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.

### 7.3.4 Prediction trend chart (C4)

The prediction trend chart empowers care providers to observe the evolution of prediction values. They can customize the timeframe to assess the development over an extended period. 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.

### 7.3.5 Patient information (C5)

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 both 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/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.

# 8 Evaluation

The evaluation focuses on a task-based think-aloud study conducted to assess the high-fidelity design, which was developed in the previous step. The qualitative nature of this study aims to measure the impact of the designed dashboard on trust & reliance, usability and decision-making.

## 8.1 Methodology

This section provides an overview of the methods used, including participants, materials and the procedure. The main method used in this study is a task-based think aloud study which is widely recognised method in usability research [45]. Participants are asked to articulate their thoughts aloud while performing tasks, providing important insights into their decision-making processes. Six distinct tasks emulated real-world scenarios, two of those tasks from another study were incorporated to evaluate the prediction model without a dashboard.

### 8.1.1 Participants

In total 7 caregivers (see Table 3) took part in the evaluation of the high-fidelity prototype, actively working in the Neonatal Intensive Care Unit (NICU) at Utrecht's Wilhelmina Children's Hospital. The gender distribution within the sample consisted of six females and one male. The sample as a whole was relatively average (M = 47.57, SD = 11.67). This group consisted of various roles within the NICU, including neonatologists, NICU physicians assistants, neonatologist-fellows, and NICU nurses. The participants exhibited diversity not only in their professional roles but also in terms of practical experience (M = 12.5 years, SD = 7.99 years). Due to limited availability, three caregivers who had previously participated in the low-fidelity interview sessions were also included in the evaluation. Because there was over 2 months between the low-fidelity prototype and evaluation session, most of the details had already been forgotten.

| Participant | Position (full) | Position (D/N) |
|---|---|---|
| P1 | Physician assistant | D |
| P2 | Nurse | N |
| P3 | Neonatologist | D |
| P4 | Fellow | D |
| P5 | Physician assistant | D |
| P6 | Physician assistant | D |
| P7 | Nurse | N |

Table 3: In total 7 caregivers of various roles took part within the NICU. The table also shows which role is considered a doctor (D) and which role is considered a nurse (N).

### 8.1.2 Material

Participants were handed a computer with the developed XAI-interface already opened so that they could perform the tasks. Interviews were recorded using a cell phone and screen recordings were captured using pre-installed Apple software. Care providers scheduled their own private rooms, so that they could take part in the evaluation without interruption. SUS-questionnaires and informed consent forms were printed. The evaluation assesses three primary variables: usability & comprehension, trust & reliability, and decision-making. Furtermore, software used to analyse the data included Amberscript to (partly) automatically transcribe audio-recordings to text and NVIVO was used to code text and extract themes. The patient data used in the dashboard was extracted from the EHR, anonymized and shuffled.

**Trust & Reliability** The effect of contextual information on the trust and reliance of predictions will be qualitatively evaluated by combining component analysis and semi-structured interviews. To evaluate the reliance, it will be investigated which contextual information has the most influence on the interpretability of a prediction by performing a component analysis. Here, contextual information is broken down into components such as patient information, vital signs, and feature contributions. During the evaluation, a sample prediction will be displayed and asked which components are most critical for interpreting this prediction. In addition, a semi-structured interview should evaluate the effect of contextual information on users' trust of predictions.

### 8.1.3 Usability

The System Usability Scale (SUS) questionnaire was used to measure user satisfaction, as detailed in Appendix D. The SUS is a standardized questionnaire

used to gauge the usability of an interface. Comprising 10 questions, participants express their agreement levels on a Likert scale. 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. A higher SUS score correlates with enhanced usage and greater user satisfaction.

**Comprehensibility** Qualitative evaluation through semi-structured interviews delved into comprehensibility of the designed components.

### 8.1.4 Task-based think-aloud analysis

To evaluate the effectiveness of the designed high-fidelity prototype regarding usability & comprehensibility, reliability & trust, and decision-making, six tasks were formulated. Task 5 and 6 were part of a parallel study conducted by a student-colleague, investigating how providers respond to prediction models without a dashboard with contextual information, as will be described in section 8.1.5. In order to minimize the impact of learning-effect, the sequence of tasks were rearranged for participants. For each participant, the two studies were swapped. Furthermore, the order of tasks 1 to and including 4 were shuffled using the Latin square method. For instance, the first participant competed tasks 1 through 6 in that specific order, while the second participant started with tasks 5 and 6, followed by tasks 2, 3, 4, and 1. Every task included a distinct prediction type (e.g., true positive, false positive, etc.), stated below within brackets, which was deliberately withheld from the participants.

**Task 1: An alarm was raised at 13:00 for a patient at high risk of sepsis (true positive). 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: An alarm was raised at 17:00 for a patient at high risk of sepsis (false positive). 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: No alarm was generated for the following patient (true negative). 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: No alarm was generated for the following patient (false negative). 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 on the dashboard, and more importantly how they arrive at this conclusion.

**Task 5: Read scenario 1 (see Appendix F). Are you concerned and how do you estimate the risk of sepsis? You may also retrieve information from the electronic patient record, look at the monitor or glance at the patient.** *Rationale task:* find out how healthcare providers make decisions without a dashboard including contextual information, and how this compares with estimating risk with a dashboard.

**Task 6: Read scenario 2 (see Appendix F). Which of these patients are you most concerned about? Which of these patients would you prioritize right now? You may also retrieve information from the electronic patient record, look at the monitor or glance at the patient.** *Rationale task:* find out how care providers prioritize patients and how decisions are made when multiple alarms are generated without a dashboard including contextual information.

### 8.1.5 Procedure

The session started with an introduction of the researcher, an explanation of the research goal and what is expected from them 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. 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 dashboard and participants could ask any questions during the explanation. Then, participants were asked to take place behind the computer, as detailed in section 8.1.4. Before they started with the first task, participants were asked to think-aloud while doing the tasks: what information they were looking at, what they were doing, what they liked and didn't 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 to them for estimating the risk. After tasks 1 until and including 4 were completed, the participant 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 can be found in the appendix G.

**Parallel study** Together with a colleague-student, a parallel study was conducted. The purpose of this study was to investigate how healthcare providers responded to ML-based systems in clinical practice, including determining urgency. Participants were asked to read scenarios from task 5 and 6 (see Appendix F) aloud to then estimate the patient's risk, as detailed in 8.1.4. They could do this by using the introduction text, a screen shot of the monitor, a screen shot of the electronic health record, a picture of the patient and the risk analysis including the final risk score. The risk analysis including the risk score was similar to the risk analysis presented on the dashboard (Screenshot is available in the appendix F). This was different from our study, were contextual information was presented in one central overview. After the tasks were completed, participants were asked whether they believed the dashboard supported interpreting risk scores or whether they preferred using existing information systems such as the Electronic Health Record (EHR).

### 8.1.6 Analysis

Data collected from the SUS was analyzed quantitatively and the scores were calculated, as detailed in 8.1.3. After transcribing the interviews and watching back the screen recordings, the emerging themes regarding the impact of contextual information on trust & reliability and decision-making were carefully examined.

## 8.2 Results

This section will present qualitative insights obtained from task-based think-aloud sessions with healthcare providers for each of the variables: usability and comprehensibility, trust and reliability, and decision-making, with each of them containing subthemes, as detailed in the following Table 4.

| **Thematic area** | **Subtheme** |
|---|---|
| Trust & Reliance | Assessment of clinical relevance |
| | Alignment with domain knowledge |
| | Perceptions of predictive performance |
| | |
| Decision-making | Risk interpretation |
| | Influence of predictions on decision-making |
| | |
| Comprehensibility & Usability | Centralized information display |
| | Interpretability explanations |
| | Further improvements |

Table 4: Overview of the themes, with corresponding subthemes, extracted from the task-based think-aloud study and semi-structured interviews acquired during the evaluation of the high-fidelity prototype.

### 8.2.1 Trust & reliance

Establishing trust in a predictive tool is paramount for decision-making and successful adoption in clinical practice. This section delves into the multifaceted aspects of trust and reliance, exploring key sub themes that emerged during the evaluation. These include the assessment of clinical relevance, alignment with domain knowledge, and perceptions of predictive performance.

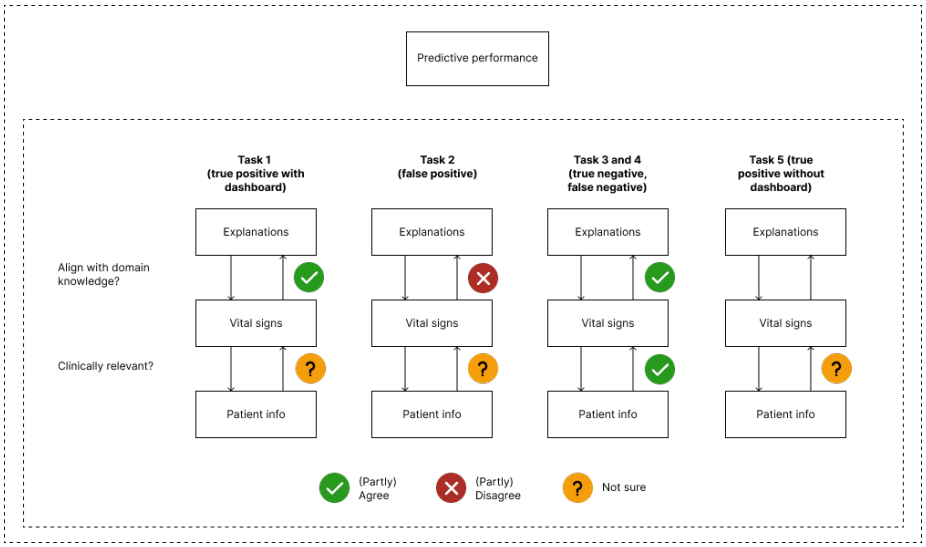



Figure 24: The presented figure illustrates the interplay between care providers' trust in the predictive model and their assessments of clinical relevance and alignment with domain knowledge across different prediction scenarios. The analysis is centered around four distinct tasks: true positive, false positive, true negative, and false negative, with a final task featuring a scenario without a dashboard. In the true positive task, care providers inspected predictions by aligning domain knowledge through comparing risk analysis and feature contributions with the vital sign chart. However, determining clinical relevance posed a challenge due to missing, unreliable, or outdated contextual information, impacting their trust in the prediction. Conversely, in the false positive task, providers found discrepancies between the explanations and vital signs, resulting in a lack of trust. In the true negative and false negative cases most participants agreed with the predictions. The alignment with domain knowledge and the establishment of clinical relevance were straightforward, as fewer incidents required less clinical context for understanding. In the final task, absent of a dashboard, providers did not check whether the predictions aligned with domain knowledge.

**Assessment of clinical relevance** When we asked participants to assess the risk of a patient and whether they agreed with the prediction, they all relied heavily on contextual patient information. 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. We observed that participants first analysed 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 at 25 weeks, now four weeks old and a little rumbling saturation*

*I don't think is a big deal"*. Another care provider (P1) completely changed it's mind after discovering that general patient information was included in the interface: *"Oh, but you have here what it is. Oh wait a minute, I don't pay attention. 25 weeks, which is now 26 days old. 860 gram. no, no then I'm not so very worried"*. Providers also relied heavily on actual interventions (e.g., respiratory support), medication (e.g., antibiotics) to determine patient baselines. For instance, participant (P2) noted: *"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. This is indicated by a question mark in figure 24. Participant 3 (P3) responded to the inquiry about her concern for the patient in task 1 with the following statement: *"So I would then indeed want to know: is there support start, yes or no? Or did they already have it? And is it intensified yes or no? And is that pressure being administered properly?"* and desired more information on previous infection episodes: *"A child who's already had an NEC once or who has an infected trhombos. You know, then the story becomes different for me anyway"*. 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: *"I do need to know if it is real (meaning: not due to interference e.g., bathing, nursing, feeding), I would check. 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 providers 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 as a crucial factor influencing care providers' trust in the model. Nevertheless, the available data on the dashboard proved insufficient for a comprehensive assessment.

**Alignment with domain knowledge** 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 domain knowledge. When predictions aligned, such as in the true positive case, trust in the model increased (see Figure 24). Participant 6 (P6) noted: *"Because there was really a huge increase in the number of bradycardias that wasn't there before. So I do understand that that's getting into a higher risk area"*. However, this was also true the other way around: trust in the model decreased when predictions did not align with clinical knowlege. For instance, participant 3 (P3) noted that the prediction trend increased, while the vital signs stabilized: "Why does it alarm at five? Because here you actually have a very stable heart action again". Interestingly, all contextual information in one central overview, did support participants in determining whether the prediction aligned with clinical domain knowledge, as indicated with the absence of the an icon for task 5 in Figure 24.

Thus, they compared the time when an alarm was generated and the progression of predictions with the vital signs to see if this corresponded to domain knowledge, which occurred to a lesser extent in the case without contextual patient information in one centralized overview.

**Perceptions of predictive performance** The perceptions of predictive performance also influenced trust. For instance, participant 2 (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 infection, then I'll definitely take that on board.". These findings were confirmed by other participants, for example, participant 3 (P3) mentioned: *"To what extent does this help me (referring to the prediction model), I know that from the monitor trend a little bit more. You know that, you have that feeling, you've been doing that for years like that"*. Participant 4 (P4) wanted to know how accurate the model was first: *"I would also like to know how that has been in previous predictions, whether it was often wrong"*. Moreover, trust in the predictions was also impacted by the limitations of the model. Participant 3 (P3) reacted surprised when she saw (through the feature importance chart) that only saturation and heart frequency were used: *"Based only on heart rate and saturation doesn't tell me very much"*. Another participant (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"*. And care providers thought that one prediction per hour was not enough, as participant 4 (P4) mentioned: *"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. Participant 4 (P4): *"I think it's that you just that at least sometimes I'm looking or looking for additional confirmation that it's high-risk."*.

### 8.2.2 Decision-making

In this section, we will delve into the outcomes concerning decision-making, exploring the two sub themes: risk interpretation and the influence of predictions on decision-making.

**Risk interpretation** Similar to findings from the low-fidelity prototype, care providers expressed a need for clarity and meaning in the results. They sought predefined rules for interpreting risks, with participant 4 (P4) highlighting the desire for a specific cut-off value. P4 articulated, *"it is also good that you then look with your group of how are we going to use this then and where do you put a cut-off value."*. Another participant (P3), also faced challenges in interpretation, questioning, *"how are you going to interpret it? Are you going to say high-risk then by definition you have to start my antibiotics."*. P3 also struggled to gauge the degree of risk within a given range, stating, *"Because still with high-risk you have a kind of gray area of, is this 85% or 99% chance*

*that it could be an infection?"*. This underscores a prevalent need for guidance on utilizing the model and interpreting risk predictions.

**Influence of predictions on decision-making** Participant 4 (P4) noted that when the prediction diverged from their own assessment, as observed in task 2, a 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 they really are declines 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) one 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 explanations to understand why the prediction deemed the patient high-risk, comparing it with vital signs and engaging in reasoning, questioning whether there was genuinely no saturation drop. This suggests that high-risk predictions have a noticeable impact on care providers, prompting them to conduct a more thorough examination of the patient and reevaluate their initial assessments.

### 8.2.3 Usability

In this section, we will explore the findings from the SUS questionnaire and delve into the qualitative results pertaining to three subthemes: centralized information display, interpretability explanations, and potential areas for further improvement.

**SUS-questionnaire** The resulting SUS-scores are presented in Table 5, with an average SUS-score of 86.25. The score was marginal for two participants (P3 and P4), and excellent for the other participants (P1, P2, P5, and P6).

| 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 |

Table 5: Overview of results from the System Usability Scale (SUS) questionnaire

**Centralized information display** Participants found it convenient that relevant information was centralized in one central overview. Participant 1 (P1) expressed this sentiment, stating, *"In Metavision I have to click through screens and then I don't have that next to it. So I like that I immediately know oh, yes, there is a line in it, from so many days. And that you see last recent lab with the time added."*. This sentiment was echoed by other participants, with participant 4 (P4) noting, *""It is nice that you have it at a glance. I do like it in itself because in Metavision you have to press different tabs all the time, and here you have everything in one overview"*. The centralized view offered a clear overview, aiding care providers in verifying risk predictions and quickly assessing patients, particularly beneficial for those commencing their shifts.

Moreover, the centralized overview proved advantageous for presenting cases to a doctor, as noted by participant 2 (P2): *"That certainly means that I get my overall picture clearly and can then alert the doctor"*.

While the centralized overview supported participants in verifying risk predictions and quickly assessing patients, concerns arose about the feasibility of maintaining this central location when missing contextual patient information was included (e.g., patient history, additional timelines). Participant 3 (P3) expressed doubt, stating, *"Well, I think you just need more patient information,*

*so to speak, and the question is whether you should put it all in that dashboard.".* Participants 4 (P4) and 5 (P5) suggested that the explanations, encompassing the risk analysis and feature importance chart, could also be integrated into the electronic health record (EHR) alongside vital signs. This integration would ensure the availability of all pertinent information while preserving the advantages offered by the explanations. In summary, participants consistently preferred the dashboard with centralized contextual information over the version without it. Nevertheless, their information needs remained unfulfilled, indicating uncertainty regarding the added value.

**Interpretability explanations** Overall, participants reacted very positive to the explanations, including the risk analysis and feature importance plot. Throughout the tasks, they consistently referred to these explanations, finding it straightforward to comprehend the evolution of prediction trends and discern features influencing the prediction. Although participant 1 (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.

**Further improvements** While the majority of participants expressed positive feedback for the dashboard featuring contextual information in one central location, they also provided suggestions for improvement. Specifically, participants, including P1, P3, P4, and P6, proposed that the vital sign chart would be more valuable for validating predictions if it exclusively displayed validated signs. This entails the removal of disturbances like feeding and nursing that may influence vital signs, thereby presenting only real validated signs. Such refinement would streamline the validation process, allowing for quick and accurate risk assessment without the need for additional checks.

Additionally, participants expressed the need for the capability to zoom in on the vital sign charts to discern patient baselines, aiding in understanding whether certain patterns, such as saturation drops, were normal for a particular patient. The integration of vital sign counts alongside the charts was deemed helpful, eliminating the need for manual counting. However, participants stressed the importance of establishing clear thresholds for what constitutes as a count, aligning with the criteria utilized in the electronic health record.

Participants also identified areas for improvement in the dashboard's design. Some overlooked the demographic patient information crucial for establishing baselines, emphasizing the need for greater visual prominence. Moreover, participants advocated for the removal of outdated or irrelevant parameters from the dashboard to maintain its relevance and utility.

# 9 Discussion

This study aimed to design an XAI interface with contextual patient information in one central location to support care providers in interpreting sepsis predictions. Addressing a recognised gap in existing XAI literature, the study focused on providing relevant contextual information alongside machine learning predictions to enhance the assessment and trust in model predictions. This aligns with broader HCI research goals of creating trustful and purposeful interfaces for ML-based tools in clinical practice, a crucial factor for adoption.

The study revealed positive outcomes regarding the usability and comprehensibility of the designed interface. Participants consistently preferred the dashboard with contextual patient information in a central location over the existing approach. However, concerns were raised about missing or outdated information, affecting participants' ability to assess the clinical relevance of predictions and overall patient risk.

## 9.1 Research Question 1.1: Specific Contextual Information for Interpretation

Providers required a diverse range of contextual information to interpret sepsis predictions. Vital signs emerged as a crucial component to interpret predictions, assessing whether the severity of detected abnormalities aligned with the severity of the risk score. This assessment was pivotal in determining the level of trust in predictions. Additionally, they utilized vital signs to gauge clinical relevance, establishing baseline values for patients by looking back into the patient's vital sign history. This finding aligns with the results of Barda et al. [6], wherein they identified the significance of time-series plots for identifying anomalous vital sign values and establishing patient baselines to assess the clinical relevance of ML predictions. This might highlight potential challenges for models with numerous dynamic parameters, as displaying all vital signs in one centralized overview may result in information overload. Providers found aggregated data, like vital sign counts, beneficial, suggesting a possible solution. Using aggregated data could help emphasize suspicious values, identify baselines, and reveal trends in a more efficient manner.

Moreover, contextual information such as patient demographics and actual interventions and medications were also important to caregivers. However, clinical information such as mobility and skin color posed challenges in integration, with concerns about reliability and the need of physically assessing patients. Also lab results were perceived as less helpful, often considered outdated. This indicates that key contextual information should be prioritized, and suggests that the design should find innovative ways to incorporate clinical data such as mobility and skin color.

Participants expressed positive reactions towards explanations, finding them valuable in assessing predictions. This indicated that the feature importance chart, risk analysis, and vital signs combined may be a viable approach in explaining predictions in clinical practice.

## 9.2 Research Question 1.2: Presentation of Contextual Patient Information

The interface's design was well received by care providers, indicating the advantage of having all information in one overview compared to the electronic health record. Providers found the dashboard easy to use, aligning with the intended simplicity. Moreover, they favoured the risk score and risk score analysis with three color-coded risk categories for intuitive interpretation. This observation aligns with the findings of Sandhu et al. [9], who noted that simplifying the visual display of sepsis risk into three colored categories reduced cognitive burden.

While the consolidated view supported validating model predictions, challenges remain in determining if all relevant information can be effectively presented in one central location. Data aggregation and showing additional details on demand may offer a comprehensive overview. However, when multiple prediction models will be introduced in the NICU, integration strategies become imperative. This integration should aim to make all contextual information readily available for quick interpretation without disrupting the workflow-a crucial factor in the adoption of systems in clinical practice [9].

Moreover, there was a clear disparity between the model's intended use and how care providers perceived its application. Caregivers sought to correlate actions with risk predictions, such as initiating treatment for high-risk patients, contrary to the model's intended purpose of alerting caregivers to potential risks for patient check-ins. Aligning with this, interface design can play a crucial role in shaping usage and operational modes. For instance, adjusting risk categories to signify alarms only or no alarm could better align with the model's intended function. Clear communication to caregivers about the necessity of validating alarms by physically assessing the patient, possibly through accompanying text, could reinforce the desired mode of operation.

## 9.3 Research Question 1.3: Impact on Trust & Reliance, Usability and Decision-Making

Our study, consistent with Matthiesen et al. [10], revealed that care providers mostly relied on their established methods to assess patient risk, treating the model as a confirmation or second opinion. Initial skepticism regarding the model's predictive performance was evident, with providers expressing a desire to understand the model's performance. While presenting detailed metrics like specificity and sensitivity may be challenging for care providers to comprehend, the absence of such metrics led to doubts about the model's predictive capabilities. This underscores the importance of intuitive methods to communicate prediction confidence, such as a color-coded confidence level display or user-friendly language.

Moreover, perceived limitations of the model, such as the number of predictors, had an adverse impact on trust, resonating with insights from Barda et al. [6]. Simultaneously, providers expressed interest in a system like the HERO

system, which relies solely on heart rate for sepsis prediction. This suggests a potential gap in understanding machine learning concepts. To address this, incorporating an overview section in the interface could potentially clarify the model's capabilities and limitations.

## 9.4 Limitations

This section outlines the limitations inherent in this study, which provide context for interpreting findings.

Firstly, caregivers unfamiliar with patients needed additional contextual information to interpret predictions, unlike those who had long-term relationships with the patients. The NICU ward, where the model operates, holds crucial information not available in the room, such as real-time patient conditions (e.g., skin color, mobility) and visible interventions (e.g., lines and respiratory support). Consequently, the contextual information required for a dashboard may vary in a real-setting study.

Additionally, constraints prevented the use of real patient information in this study, leading us to employ anonymized data. The absence of pseudonymized patient information for evaluation led to the shuffling of general patient details, occasionally resulting in notable cases. Despite potential differences in choices without shuffled information, the impact on the study's outcomes related to contextual information interpretation, comprehensibility, and verification remained minimal.

Finally, participants overlapped between the design and evaluation phases, where some had seen parts of the dashboard earlier, possessing additional knowledge about the model's limitations. Due to staffing constraints and a busy clinic, we had to approach available participants, including those 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.

## 9.5 Future work

In this section, future research will be discussed which can contribute to the refinement of predictive model interfaces in healthcare settings.

Firstly, 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.

Additionally, 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.

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.

Integrating multiple predictive models into a unified dashboard or the EHR is another area that needs exploration. Research should investigate how to present information from diverse models in a one consolidated overview, allowing care providers to interpret and act upon predictions without disrupting clinical workflows.

Furthermore, the study suggested that data aggregation could play a crucial role in providing a comprehensive overview of all relevant contextual information. Future work should look into ways on how to aggregate data, considering visualization techniques.

# 10 Conclusion

In conclusion, our study aimed to design an explainable interface to support care providers in interpreting and validating sepsis risk predictions. Throughout this study, we delved into specific care provider needs, uncovering key insights about the presentation of contextual information, and the impact on trust & reliance, usability and decision-making.

## 10.1 RQ1.1: What specific contextual information do healthcare professionals require to interpret sepsis predictions?

Addressing the first research question (RQ1.1), we found that vital signs, risk analysis, and feature importance charts supported care providers in aligning predictions with domain knowledge. Although contextual patient information was necessary to assess the clinical relevance of predictions, we found that contextual information on the dashboard was deemed insufficient, unreliable, or not suited for real-time patient assessment.

## 10.2 RQ1.2: How can contextual information be best presented in an XAI interface to facilitate the interpretation of sepsis predictions?

Moving on to question (RQ1.2), our findings highlighted that explanations, such as tornado plots and risk analysis, were perceived as easy to interpret. Combining these explanations with vital signs was suggested as a viable approach. Providers favored an overview-style presentation for decision-making and risk interpretation, opting for this over the current EHR in combination with risk trends. The importance of displaying only validated vital signs, enabling zooming for trend analysis, incorporating timestamps, and addressing visual preferences across different care providers' roles were emphasized.

## 10.3 RQ1.3: What is the impact of a dashboard with contextual information on trust, reliance and decision-making for healthcare professionals?

Exploring the impact of the dashboard on trust, reliance, and decision-making (RQ1.3), we discovered that an overview in one central location was well-received and positively influenced decision-making. Validated predictions, aligning with domain knowledge, increased trust, while explanations revealing the limited number of features used in the model and a lack of proof of model accuracy decreased trust. This initial skepticism led to the model being used as a second opinion, with healthcare providers validating their assessments against the model's predictions.

# References


[1] Christina L Cifra, Jason W Custer, and James C Fackler. A research agenda for diagnostic excellence in critical care medicine. *Critical care clinics*, 38(1):141–157, 2022.

[2] Fumiaki Nakamura and Michikazu Nakai. Prediction models—why are they used or not used?—. *Circulation Journal*, 81(12):1766–1767, 2017.

[3] Amina Adadi and Mohammed Berrada. Explainable ai for healthcare: from black box to interpretable models. In *Embedded Systems and Artificial Intelligence: Proceedings of ESAI 2019, Fez, Morocco*, pages 327–337. Springer, 2020.

[4] Hubert D Zajac, Dana Li, Xiang Dai, Jonathan F Carlsen, Finn Kensing, and Tariq O Andersen. Clinician-facing ai in the wild: Taking stock of the sociotechnical challenges and opportunities for hci. *ACM Transactions on Computer-Human Interaction*, 30(2):1–39, 2023.

[5] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38, 2019.

[6] Amie J Barda, Christopher M Horvat, and Harry 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(1):1–16, 2020.

[7] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15, 2019.

[8] Zhuochen Jin, Shuyuan Cui, Shunan Guo, David Gotz, Jimeng Sun, and Nan Cao. Carepre: An intelligent clinical decision assistance system. *ACM Transactions on Computing for Healthcare*, 1(1):1–20, 2020.

[9] Sahil Sandhu, Anthony L Lin, Nathan Brajer, Jessica Sperling, William Ratliff, Armando D Bedoya, Suresh Balu, Cara O'Brien, and Mark P Sendak. Integrating a machine learning system into clinical workflows: qualitative study. *Journal of Medical Internet Research*, 22(11):e22421, 2020.

[10] Stina Matthiesen, Søren Zöga Diederichsen, Mikkel Klitzing Hartmann Hansen, Christina Villumsen, Mats Christian Højbjerg Lassen, Peter Karl Jacobsen, Niels Risum, Bo Gregers Winkel, Berit T Philbert, Jesper Hastrup Svendsen, et al. 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 human factors*, 8(4):e26964, 2021.

[11] waarschuwingssysteem-spoort-bloedvergiftiging-vroegtijdig-op-bij-premature-baby, howpublished = https://www.vakbladvroeg.nl/waarschuwingssysteem-spoort-bloedvergiftiging-vroegtijdig-op-bij-premature-baby/, note = Accessed: 2023-08-12.

[12] Nadine Bienefeld-Seall, Rahel Lüthy, Dominique Brodbeck, Jens Boss, Jan Willms, Jan Azzati, Mirco Blaser, and Emanuela Keller. Solving the explainable ai conundrum: How to bridge the gap between clinicians needs and developers goals. *Research Square*, 2022.

[13] Martin Stocker, Sina B Pilgrim, Margarita Burmester, Meredith L Allen, and Wim H Gijselaers. Interprofessional team management in pediatric critical care: some challenges and possible solutions. *Journal of multidisciplinary healthcare*, pages 47–58, 2016.

[14] Annika Kaltenhauser, Verena Rheinstädter, Andreas Butz, and Dieter 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*, pages 1509–1522, 2020.

[15] Sheila M Gephart, D Anthony Tolentino, Megan C Quinn, and Christina Wyles. Neonatal intensive care workflow analysis informing nec-zero clinical decision support design. *CIN: Computers, Informatics, Nursing*, 41(2):94–101, 2023.

[16] Minfan Zhang, Daniel Ehrmann, Mjaye Mazwi, Danny Eytan, Marzyeh Ghassemi, and Fanny 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*, pages 1–13, 2022.

[17] Tom Taulli and Michael Oni. *Artificial intelligence basics*. Springer, 2019.

[18] Gary Marcus and Ernest Davis. *Rebooting AI: Building artificial intelligence we can trust*. Vintage, 2019.

[19] Robert Munro Monarch. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster, 2021.

[20] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[21] Tianyi Li, Mihaela Vorvoreanu, Derek DeBellis, and Saleema Amershi. Assessing human-ai interaction early through factorial surveys: A study on the guidelines for human-ai interaction. *ACM Transactions on Computer-Human Interaction*, 2022.

[22] Mansoureh Maadi, Hadi Akbarzadeh Khorshidi, and Uwe Aickelin. A review on human–ai interaction in machine learning and insights for medical applications. *International journal of environmental research and public health*, 18(4):2121, 2021.

[23] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.

[24] Daniel Kahneman and Amos Tversky. *The simulation heuristic*. National Technical Information Service, 1981.

[25] Mireia Ribera and Àgata Lapedriza García. Can we do better explanations? a proposal of user-centered explainable ai. CEUR Workshop Proceedings, 2019.

[26] Gabriëlle Ras, Marcel van Gerven, and Pim Haselager. Explanation methods in deep learning: Users, values, concerns and challenges. *Explainable and interpretable models in computer vision and machine learning*, pages 19–36, 2018.

[27] Brian Y Lim, Qian Yang, Ashraf M Abdul, and Danding Wang. Why these explanations? selecting intelligibility types for explanation goals. In *IUI Workshops*, 2019.

[28] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

[29] Tianyi Zhang, Jarrod Mosier, and Vignesh Subbian. How do clinicians use electronic health records for respiratory support decisions? a qualitative study in critical care. 2023.

[30] Tamara Munzner. *Visualization analysis and design*. CRC press, 2014.

[31] Neda Rostamzadeh, Sheikh S Abdullah, and Kamran Sedig. Visual analytics for electronic health records: a review. In *Informatics*, volume 8, page 12. MDPI, 2021.

[32] David Gotz, Fei Wang, and Adam Perer. A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *Journal of biomedical informatics*, 48:148–159, 2014.

[33] Bum Chul Kwon, Janu Verma, and Adam Perer. Peekquence: Visual analytics for event sequence data. In *ACM SIGKDD 2016 Workshop on Interactive Data Exploration and Analytics*, volume 1, 2016.

[34] David Gotz and Harry Stavropoulos. Decisionflow: Visual analytics for high-dimensional temporal event sequence data. *IEEE transactions on visualization and computer graphics*, 20(12):1783–1792, 2014.

[35] Adam Perer, Fei Wang, and Jianying Hu. Mining and exploring care pathways from electronic medical records with visual analytics. *Journal of biomedical informatics*, 56:369–378, 2015.

[36] Sally L Baxter, Jeremy S Bass, and Amy M Sitapati. Barriers to implementing an artificial intelligence model for unplanned readmissions. *ACI open*, 4(02):e108–e113, 2020.

[37] Casper de Haas. Usability study of an explainable machine learning risk model for predicting illegal shipbreaking. Master's thesis, 2021.

[38] Merel AM van den Berg, OAG O'Jay, Manon MJNL Benders, Richard RT Bartels, Daniel DC Vijlbrief, et al. Development and clinical impact assessment of a machine-learning model for early prediction of late-onset sepsis. *Computers in Biology and Medicine*, page 107156, 2023.

[39] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. I think i get your point, ai! the illusion of explanatory depth in explainable ai. In *26th International Conference on Intelligent User Interfaces*, pages 307–317, 2021.

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

[41] Costin Pribeanu. A revised set of usability heuristics for the evaluation of interactive systems. *Informatica Economica*, 21(3):31, 2017.

[42]

[43] Sander Treur. Designing an interface for an explainable machine learning risk model for predicting illegal shipbreaking. Master's thesis, 2022.

[44] Kurt Koffka. *Principles of Gestalt psychology*, volume 44. Routledge, 2013.

[45] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, 2017.

# A Discover: Interview outline

## A.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?

## A.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)

## A.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?

## A.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?

# B Discover: Interview outline - information sharing

## B.1 General

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

## B.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?

## B.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?

## B.4 Collaboration

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

# C 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?

# D System Usability Scale (SUS)

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

# E Informed Consent

## E.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 real-world 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.

## E.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.

## E.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.

## E.4 Your agreement

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

# F Evaluation: Scenarios

## F.1 Scenario 1

It is evening and you are starting duty in the NICU and have just had the handover. One of the children admitted is infant Paula. Paula was born preterm at 31+2 weeks ammenorrhea after difficult last weeks of pregnancy with ruptured membranes (PPROM) 24 hours before delivery. Delivery went smoothly with Apgar scores of 6, 8 and 9 after 1, 5 and 10 minutes postpartum. Subsequently, there were mild transition problems on the first day for which a day and a half of nasal CPAP with good recovery. On life day 4, the less with Paula: she is irritable, cries a lot and is less active in between, in addition she is tachycardia and has a rectal temperature of 35.8 degrees Celsius.


Figure 25: Enter Caption

## F.2 Scenario 2

It is evening and you begin your shift in the NICU and have just had the handover. You have the following patients:

1. Paula - born at 31+2 weeks and a birth weight of 1.1 kg. Paula is irritable, cries a lot and is less active in between, in addition she is tachycardic and she has a rectal temperature of 35.8 degrees Celsius.

2. Jonathan - born at 28+3 weeks and a birth weight of 930 grams. Jonathan is less active, sleeps a lot, tachypneic (up to 80 per minute), a heart rate of 170 per minute and a temperature of 38.0 degrees Celsius.

3. Floor - born at 24+3 weeks and a birth weight of 680 grams. With Floor has been doing well for the past few days. She is sleeping well now and has a nice pink color.

4. Joep - born at 32+1 weeks and a birth weight of 1.2 kg. Also with Joep is also doing well this past week. He is sleeping well now and looks good looking.

# G Evaluation: Task-based think-aloud protocol

Demographic questions:

- Age
- Domain-expertise
- Years of experience

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 containes enough information?
- How difficult did you find the task? (cognitive load)

After all 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?