Design and Evaluation of User-Centered Explanations for Machine Learning Model Predictions in Healthcare

by

Amie Janeth Barda

Bachelor of Science, The Ohio State University, 2013

Master of Science, University of Pittsburgh, 2015

Submitted to the Graduate Faculty of the

School of Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

SCHOOL OF MEDICINE

This dissertation was presented

by

Amie Janeth Barda

It was defended on

December 12, 2019

and approved by

Dr. Michael Becich, Professor and Chair, Department of Biomedical Informatics

Dr. Christopher Horvat, Assistant Professor, Department of Pediatric Critical Care Medicine

Dr. Douglas Landsittel, Professor, Department of Biomedical Informatics

Dr. Shyam Visweswaran, Associate Professor, Department of Biomedical Informatics

Dissertation Director: Dr. Harry Hochheiser, Associate Professor, Department of Biomedical Informatics

Design and Evaluation of User-Centered Explanations for Machine Learning Model Predictions in Healthcare

Amie Janeth Barda, PhD

University of Pittsburgh, 2019

Challenges in interpreting some high-performing models present complications in applying machine learning (ML) techniques to healthcare problems. Recently, there has been rapid growth in research on model interpretability; however, approaches to explaining complex ML models are rarely informed by end-user needs and user evaluations of model interpretability are lacking, especially in healthcare. This makes it challenging to determine what explanation approaches might enable providers to understand model predictions in a comprehensible and useful way. Therefore, I aimed to utilize clinician perspectives to inform the design of explanations for ML-based prediction tools and improve the adoption of these systems in practice.

In this dissertation, I proposed a new theoretical framework for designing user-centered explanations for ML-based systems. I then utilized the framework to propose explanation designs for predictions from a pediatric in-hospital mortality risk model. I conducted focus groups with healthcare providers to obtain feedback on the proposed designs, which was used to inform the design of a user-centered explanation. The user-centered explanation was evaluated in a laboratory study to assess its effect on healthcare provider perceptions of the model and decision-making processes.

The results demonstrated that the user-centered explanation design improved provider perceptions of utilizing the predictive model in practice, but exhibited no significant effect on provider accuracy, confidence, or efficiency in making decisions. Limitations of the evaluation

study design, including a small sample size, may have affected the ability to detect an impact on decision-making. Nonetheless, the predictive model with the user-centered explanation was positively received by healthcare providers, and demonstrated a viable approach to explaining ML model predictions in healthcare. Future work is required to address the limitations of this study and further explore the potential benefits of user-centered explanation designs for predictive models in healthcare.

This work contributes a new theoretical framework for user-centered explanation design for ML-based systems that is generalizable outside the domain of healthcare. Moreover, the work provides meaningful insights into the role of model interpretability and explanation in healthcare while advancing the discussion on how to effectively communicate ML model information to healthcare providers.

Table of Contents

List of Tables

List of Figures

Preface

I am deeply grateful for the unwavering support that I have received throughout this journey. First and foremost, I would like to thank my thesis advisor, Dr. Harry Hochheiser, who provided me with valuable advice and support long before becoming my advisor. It has been a privilege to be his student, and I want to express my sincere appreciation for his guidance, patience, and encouragement throughout the completion of this dissertation. I would also like to give a special thanks to my committee member Dr. Christopher Horvat, without whom this work would not have been possible. Dr. Horvat provided access to the data used in this work, recruited all study participants, and helped advise the research process. I am grateful not only for his time and advice, but for his genuine interest in my research and in my future career path. Also, I want to thank committee members Dr. Shyam Visweswaran, who helped conceive this project and provided valuable insights; Dr. Douglas Landsittel, who helped me work through several statistical challenges; and Dr. Michael Becich, who advised me through some difficult circumstances and worked to find funding to support the completion of this work.

I gratefully acknowledge the institutions that have provided me with funding support: the National Institutes of Health and National Library of Medicine (under award 5 T15 LM007059-27) and the Department of Biomedical Informatics (DBMI) at the University of Pittsburgh.

Thank you to all the staff and administrators of DBMI and Children's Hospital of Pittsburgh who helped make this work possible. I would like to specifically thank Vickie Johnson for reserving conference rooms for study sessions. A very special thanks also goes to Toni Porterfield, who has always gone above and beyond for me. I appreciate all she does as an administrator, friend, and confidant of the students in DBMI.

I am indebted to all those who provided me with opportunities to enrich my graduate school experience. Thanks to Dr. Rich Tsui and all the members of the Tsui lab for providing a supportive and intellectually challenging research environment for the first several years of this journey. It was a privilege to know and work with you all. Thanks to Dr. David Boone and the Computer Science, Biology, and Biomedical Informatics Summer Academy for providing me with the opportunity to hone my teaching and mentorship skills. Thanks to the Jewish Healthcare Foundation for the knowledge and experience I gained in the Patient Safety Fellowship program. And a special thanks to Fourth River Solutions (4RS) for providing me with invaluable professional development and leadership opportunities.

Finally, my deepest gratitude is reserved for the endless patience, support, and encouragement of my friends and family. Thank you to all my Pittsburgh friends who provided years of support and companionship. Special thanks to my friend Jose Posada for helping me work through research problems and for providing support during challenging times. Another special thanks to my friend, Rissa Diehl, for always being willing to lend an ear and a place to stay. Thanks to my mother-in-law, Marisa Barda, for her endless positive energy and belief in me. Thanks to my sister, Katie Otte, for taking care of daily tasks to which I had no time to attend. Thanks to my wonderful dog, Rudy, for always putting a smile on my face at the end of a rough day. A very special thanks to my husband, Christopher Barda, for his endless love, encouragement, and belief in me at all times. Finally, the greatest thanks goes to my parents, Robert and Kristi Draper, for instilling in me a life-long love of learning and supporting me throughout this journey in more ways than I can count. This would have never been possible without you.

My sincerest thanks to all, including anyone I may have inadvertently omitted, for whatever your role has been in supporting my personal and professional endeavors.

List of abbreviations:

AI: artificial intelligence

AUPRC: area under the precision-recall curve

AUROC: area under the receiver operating characteristic curve

CDSS: clinical decision support system

CFS: correlation-based feature subset

CHP: Children's Hospital of Pittsburgh

CI: confidence interval

CPR: cardiopulmonary resuscitation

EHR: electronic health record

GCS: Glasgow coma scale

HCI: human computer interaction

ICD: International Classification of Diseases

ICU: intensive care unit

IG: information gain

IRB: Institutional Review Board

LIME: local interpretable model-agnostic explanations

ML: machine learning

SHAP: Shapley additive explanations

SVM: support vector machine

UTAUT: unified theory of acceptance and use of technology

WEKA: Waikato Environment for Knowledge Acquisition

1.0 Introduction

The healthcare industry is expected to follow the patterns of other information-rich industries and experience rapid growth in the use of statistical and machine learning (ML) techniques to leverage the predictive power of large data.[1–6] There are numerous publications demonstrating the high performance of ML models on complex problems in medicine, yet there is a distinct absence of these models in practical applications in medicine.[4,5,7–9] While it is possible that this absence could be due to a lack of generalizability or reproducibility of highly accurate ML models, many publications attribute the absence to a lack of interpretability, or a model's ability to explain its behavior.[3,5,6,8,10–12] Model interpretability is highly valued in medicine, as is evidenced by the long-standing use of less accurate, but comprehensible models such as logistic regression.[13–17] Moreover, with increasing societal concerns and regulations on intelligent algorithms,[6,18–20] recognition of the importance of incorporating providers and domain knowledge in modeling processes,[4,6,8,9,21–24] and provider demand for model explanations,[5,6,10,12,25] interpretability will be vital to the future success of ML models in healthcare.

In response to the demand for model interpretability in healthcare as well as other domains, research within the ML community has produced several approaches to explaining models and predictions. While these approaches are discussed in detail in section 2.1.2, the general purpose of an explanation is to answer a particular question a user may have about the model. As a simple example, consider a linear regression model with 100 features. One user may want to know the relationships the model learned between the features and the outcome of interest. For this user, an explanation might consist of the list of weights the model assigned to all 100 features. On the other hand, another user may only want to know why the model made a specific prediction. In this case,

an explanation might consist of the 10 features most responsible for the specific prediction. Depending on the question and the user, different explanations about the model may be required.

Recently, there has been increased attention to the apparent lack of end-user involvement in the design and evaluation of explanation approaches, despite the acknowledgement that user goals, expertise, and time constraints are central in defining explanation needs.[11,12,26–32] The definition of what constitutes a "good" or "useful" explanation is often left to the judgment of novice and expert model developers, whose knowledge and backgrounds are generally not representative of end-user expertise.[27,29,33] More specifically, most developers are mainly concerned with the statistical and modeling challenges of generating an explanation; the display of the explanation often receives less attention and is rarely informed by end-user needs or insights from the literature.[20,27,31,33] Moreover, it is unclear from current evaluation studies how end-users interpret and utilize explanations designed by modeling experts,[12] which often require some level of understanding of ML models to accurately interpret. This may lead to a lack of usability and practical interpretability of these explanations for real end-users.

In healthcare, most ML models are proposed as tools to help healthcare providers analyze patient data and derive insights that can assist in clinical decision-making.[4,12,29,34] More specifically, ML-based clinical decision support systems (CDSS) usually aim to help healthcare providers make more accurate decisions, be more confident in their decisions, and/or be more efficient in making decisions. Explanations for ML models can assist in this process by providing additional information about the model that allows the provider to integrate model information with their knowledge in order to make informed decisions. This implies that explanations must be designed to fit healthcare provider information needs. Unfortunately, there is a sparsity of interpretability evaluation focused on medical applications and most claims regarding model

interpretability lack rigorous evaluations utilizing real end-users.[28,29,31,32] This makes it difficult to determine when explanations for ML models may be required and how to design these explanations to fit the information needs and environment of healthcare providers.

In this dissertation, I aimed to utilize clinician perspectives to inform the design of explanations for ML-based prediction tools. More specifically, I aimed to utilize literature insights to develop a theoretical framework of explanation design that would account for healthcare provider explanation needs when utilizing a predictive model in clinical practice. I then aimed to utilize the framework to suggest possible explanation designs that could be augmented with feedback from healthcare providers to inform the design of a user-centered explanation.

1.1 Hypotheses and Specific Aims

I hypothesized that a user-centered explanation design for an ML-based prediction tool would:

1) *Improve provider perceptions of utilizing an ML-based prediction tool in clinical practice relative to the same tool without explanations*
2) *Improve provider accuracy, confidence, and speed in making decisions relative to the same tool without explanations and having no available tool*

To evaluate this hypothesis, the following specific aims were addressed:

Aim 1. Develop a theoretical framework of explanation design and use the framework to suggest explanation designs. Using insights from the literature, develop a theoretical framework of clinical explanation design that accounts for healthcare provider explanation

needs when using a model in clinical practice. Then, use the framework to suggest explanation designs.

Aim 2. Refine explanation design with healthcare provider feedback. Conduct user studies with healthcare providers to refine explanation needs, identify successful design elements, and inform the final design of a user-centered explanation.

Aim 3. Evaluate the impact of the user-centered explanation. Conduct a laboratory study with healthcare providers to assess the impact of the user-centered explanation design on decision-making and perceptions of an ML-based prediction tool.

To ground the work in a specific context, I focused on designing and evaluating a user-centered explanation for an in-hospital mortality risk model for pediatric intensive care unit (ICU) patients.

1.2 Motivation

An ML-based CDSS can only be successful if healthcare providers accept and use the system.[35–37] System acceptance and use will likely be determined by some combination of contextual factors and design factors. To understand how a user-centered explanation may influence successful implementation of an ML-based CDSS, it is necessary to understand the specific factors that might influence system acceptance and use. The unified theory of acceptance and use of technology (UTAUT)[38] provides one possible approach to understanding such factors. The UTAUT is a validated theory of technology adoption and use that has been shown to explain some of the variance in the use of health information systems.[39] The theory, shown graphically in Figure 1, identifies performance expectancy, effort expectancy, social influence, and facilitating conditions as the four key constructs that determine user acceptance and usage behavior either

directly or as determinants of the behavioral intention to use a technology. Gender, age, experience and voluntariness of user are included as moderators of the impact of the four key constructs on usage intention and behavior. By examining how explanation design may affect key UTAUT constructs, one can obtain a better understanding of the potential role user-centered explanations may play in healthcare provider acceptance and use of an ML-based CDSS in practice. The goal in this work was to look at design issues that may impact performance expectancy and effort expectancy, as the social influence and facilitating conditions constructs would be challenging to control. These constructs deal with an individual's perceptions of how the specific social environment and organizational infrastructure promotes the use of a system, which makes it difficult to theorize how explanations might affect these constructs.

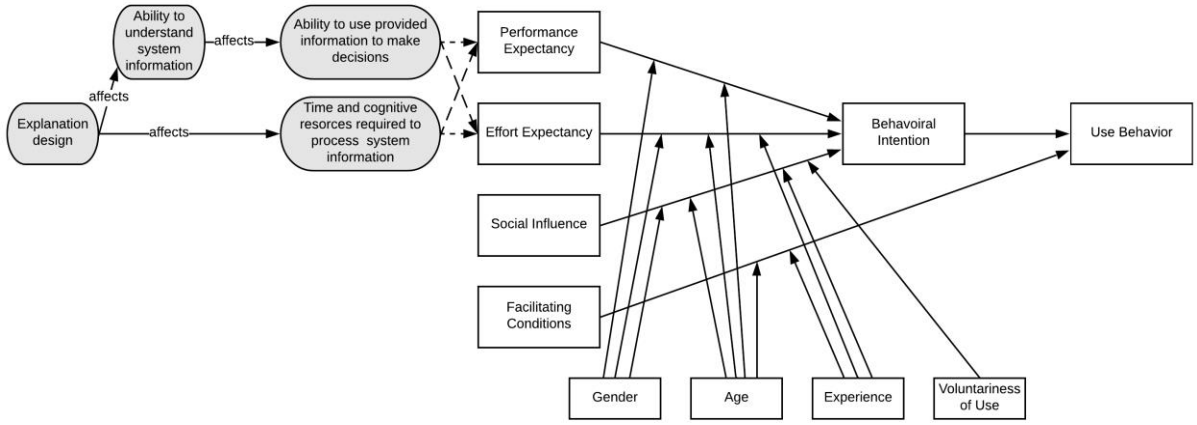


Figure 1. Graphic depiction of how explanation design may impact key constructs in the unified theory of acceptance and use of technology (UTAUT). White boxes and solid line arrows depict the main constructs and modifiers that comprise the UTAUT theory. The grey ovals indicate the proposed extensions to the UTAUT model to demonstrate the potential impact of explanation design on key constructs. The exentions are connected to their main constructs using dashed arrows. (Adapted from Venkatesh et al.[38])

Figure 1 depicts how an explanation design may affect the remaining two constructs, performance expectancy and effort expectancy. Performance expectancy refers to an individual's perceptions of how a system might improve or detract from his/her ability to do his/her job, while effort expectancy refers to an individual's perceptions of the degree of effort involved in understanding and using a system.[38] These perceptions can be directly influenced by the degree to which explanation designs for ML-based CDSS fit healthcare providers' information needs and environment. For example, how well an explanation design fits a healthcare provider's information needs likely affects the degree to which a healthcare provider can understand the information being provided by the ML-based CDSS. This directly affects whether healthcare providers can integrate the system knowledge with their own in order to make informed decisions, which can affect provider perceptions of the effort required to use the system (effort expectancy) and the utility of the system to improve their job (performance expectancy). Additionally, the degree to which explanation design fits the environment in which an ML-based CDSS is being used likely affects the time and cognitive resources required to understand the information being provided by the system. Demands on time and cognitive resources will also likely affect provider perceptions of the effort required to use the system (effort expectancy) and the utility of the system to improve their job (performance expectancy). Thus, the degree to which explanation designs for ML-based CDSS fit healthcare providers' information needs and environment may influence perceptions of the performance expectancy and effort expectancy of the system, both of which influence the behavioral intention to use the system.

Based on the UTAUT constructs and the proposed extensions, it appears that user-centered explanations for ML-based CDSS could play an important role in healthcare provider acceptance and use of the system in practice. This suggests that employing user-centered approaches to

explanation design would be beneficial. Researchers in the human computer interaction (HCI) and ML communities have proposed frameworks for and provided guidance on user-centered explanation design for systems based on ML models and other artificial intelligence (AI) approaches.[20,40–44] When discussing the design of explanations, this body of literature focuses mainly on *who* an explanation is provided to, or the user of the system, and *why* the user requires an explanation, or the specific goals the user is trying to accomplish.[40,42,44] These are important elements in understanding the context of use of an explanation, yet little attention seems to be paid to *where* or *when* users require explanations. These additional questions relate to the environment in which a user is expected to use an explanation, which has been demonstrated to be an important consideration when designing explanations for ML-based CDSS. Thus, it appears that current frameworks for user-centered explanation design do not properly account for healthcare provider explanation needs when utilizing a predictive model in clinical practice.

1.3 Approach

Combining insights from and expanding upon prior theory-informed frameworks for user-centered explanation design, I proposed a new framework for designing user-centered explanations for ML-based systems for healthcare in which explanation design is informed by the *entire* context of use. Specifically, I proposed that user-centered explanation design in healthcare should not only consider *who* an explanation is being provided to and *why* they desire that explanation, but also *when* and *where* that explanation will be used (i.e., the environment of use). The proposed framework supports explanation design by linking the components of the context of use (*who, why, when, where*) to explanation design choices such as *what* information the explanation needs to

contain (i.e., the content) and *how* that information needs to be provided (i.e., the presentation). I subsequently demonstrated an application of the framework by designing and evaluating explanations for a pediatric ICU in-hospital mortality risk model. More specifically, I used literature insights to define the context of use for the prediction tool and suggest possible explanation designs. Feedback from healthcare providers was then used to refine the defined context of use and inform the final design of a user-centered explanation. The impact of the user-centered explanation on healthcare provider decision-making and perceptions of the prediction tool was then evaluated in a laboratory study.

1.4 Significance and Innovation

To the best of my knowledge, this is the first proposed framework for user-centered explanation design for ML-based systems that provides specific guidance on design choices based on the entire context of use of the explanation. While the framework was developed and applied with a focus on explaining ML-based systems in healthcare, it is generalizable to other domains as well.

This work also provides meaningful contributions to the discussion on the importance of model interpretability in healthcare. As mentioned previously, several papers have pointed out lack of model interpretability as a barrier to the adoption of ML models in practical clinical applications.[3,5,6,8,10–12] However, a lack of model interpretability is not the only possible barrier to adoption. Other work has identified barriers that relate to model utility, such as a poor match between model information and clinical information needs (e.g., models that don't predict events of clinical relevance or that do not provide actionable information).[45,46] While model

interpretability may improve model utility (e.g., providing information that leads to actionable insights), it is unclear how the relationship between these two concepts may influence the adoption of a predictive model in practice. Moreover, it is unclear if one concept might play a more influential role in adoption. Healthcare provider assessments are needed to identify which specific factors related to interpretability and utility might prevent an ML model from being used in practice. This study is among the few that have evaluated predictive model explanations using healthcare providers, and sheds light on how user-centered explanation designs enable healthcare providers to understand an ML model in a meaningful way. Findings from this work help partially elucidate factors related to interpretability and utility that may impact the acceptance and use of ML-based tools in practice.

Finally, this work contributes to knowledge on how to communicate model predictions, specifically those based on complex ML models, to healthcare providers in a manner that facilitates their involvement in conversations about the development, deployment, and continuous improvement of predictive models for use in clinical practice. These conversations help ensure the development of ML-based systems that deliver information when and where it is needed in a way that is useful to providers and which may promote positive changes in clinical practice.

1.5 Dissertation Overview

In Chapter 2, I present an overview of the literature on interpretability, review current frameworks and guidance on user-centered explanation design for AI systems, and discuss prior work on interpretability in healthcare. Chapter 3 presents the new proposed framework for user-centered explanation design for ML-based systems in healthcare, while Chapters 4, 5, and 6

demonstrate an application of the framework. Chapter 4 demonstrates how the proposed framework was used in conjunction with literature insights to define a context of use for an ML-based prediction tool and suggest possible user-centered explanation designs. Chapter 5 describes the user studies conducted with healthcare providers to refine the context of use and explanation designs to develop a final, user-centered explanation design for the ML-based prediction tool. Chapter 6 presents an evaluation of the impact of the user-centered explanation design on healthcare provider decision-making and perceptions of the ML-based prediction tool. Finally, Chapter 7 includes a discussion of the work completed, identifies limitations, suggests directions for future work, and presents final conclusions.

2.0 Background

This chapter provides a review of the literature relevant to this dissertation. Section 2.1 provides an overview of the literature on interpretability, including how the concept is defined and its relationship to explanation, approaches to achieving interpretability, and approaches to evaluating interpretability. Section 2.2 summarizes available frameworks and guidance for user-centered explanation design and evaluation. Section 2.3 concludes the chapter with an overview of the role of interpretability in healthcare and summarizes prior work in the area.

2.1 Landscape of Interpretability

Interpretability is a multi-dimensional concept that is closely related to the concept of explanation. The reviewed literature lacks consistent terminology and definitions for these terms, which can make interpretation challenging. Figure 2 provides an overview of the concept of interpretability as viewed in this dissertation and serves as a visual guide for the concepts discussed in sections 2.1.1-2.1.2.

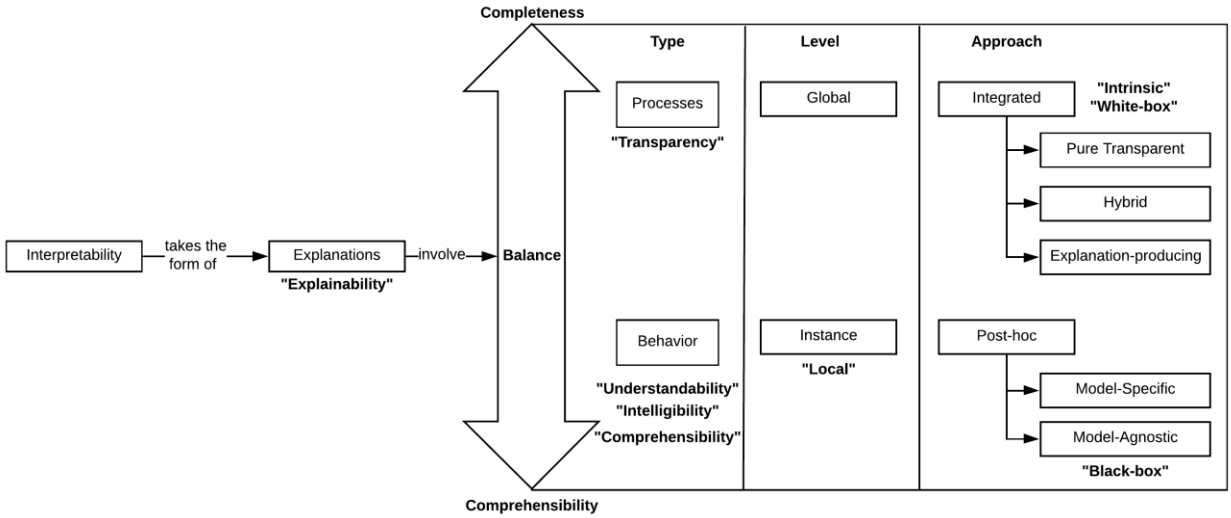


Figure 2. Landscape of interpretability. A roadmap for understanding the multi-dimensional nature of the concept of interpretability. Alternative terms commonly used in the literature are bolded and in quotes.

2.1.1 Defining Interpretability and Explanation

Current literature lacks a concrete definition for the term "interpretability"[11,30,43,47–50]; however, the demand for interpretability appears to arise when the goals of real world deployment require a system to satisfy evaluation criteria that are hard to formulize or quantify as part of the problem formulation of a system.[28,43,47] Examples of such criteria are defined in Table 1. These evaluation criteria are often interrelated and usually require subjective assessment by humans to determine if they are met. Thus, when users demand interpretability, they are often seeking some sort of explanation about a model or a system to assist them in evaluating whether certain criteria are satisfied. This close relationship between the concepts of interpretability and explanation often results in the term "explainability" being used interchangeably with "interpretability".[11,28,47,51] Interpretability generally takes the form of an explanation, but what defines an explanation and

what constitutes a quality explanation are topics still widely debated within both the ML and social science literature.[27,43,52,53]

Table 1. Examples of evaluation criteria that promote a demand for interpretability

Criteria	Definition
Fairness & Bias Reduction[6,31,43,47,49]	Ensuring that protected groups are not discriminated against
Adherence to ethical principles[6,47]	Ensuring that algorithm decisions or suggestions conform to ethical standards
Privacy[6,43]	Protecting sensitive information
Accountability & Liability[6,31,54–56]	Assigning responsibility of a suggestion or decision to an algorithm
Transferability, Reliability, & Robustness[26,43,47]	Ensuring algorithms exhibit certain levels of performance when applied in unfamiliar situations
Informativeness[31,47]	Providing useful information for real-world decision-making or accomplishing a task
Safety[28]	Protecting against danger, risk, or injury caused by decisions or suggestions of a system
Justifiability[11,21]	Ensuring a model aligns with existing domain knowledge

A basic definition of an explanation adopted in the ML literature is the concept of an 'everyday explanation', which is defined by Miller[27] as an answer to a why-question. Gilpin et al.[52] notes that this formulation of the concept of explanation is particularly interesting in ML because "when you can phrase what you want to know from an algorithm as why questions, there is a natural qualitative representation of when you have answered said question—when you can no longer keep asking why". Under this view, it can be said that the demand for interpretability is met when the ML system has provided satisfactory explanations for all questions put forth by the users of the ML system. The specific questions asked and explanations expected will depend on a user's individual relationship to the system,[56] and no single explanation is likely to satisfy all users. The differing goals, expertise (e.g., background knowledge, experiences), and time constraints of users play a central role in determining the appropriate explanation that answers a question.[12,26,43,51,52,56]

Researchers have noted the challenge of producing appropriate explanations to various users while also providing an accurate explanation of the underlying ML system

processes.[11,12,26,52,56,57] Gilpin et al.[52] describe this issue by proposing to view explanations as having properties of *comprehensibility*[1] and *completeness*. The *comprehensibility* of an explanation refers to its ability to describe a system in a way that is understandable to humans and relies on producing system descriptions that respect the cognition, knowledge, and biases of the user. In other words, the explanation must produce system descriptions that are "simple enough for a person to understand using a vocabulary that is meaningful to the user".[52] The *completeness* of an explanation is its ability to describe the operations of a system in an accurate way. These are often conflicting goals. For example, an explanation that achieves perfect *completeness* may use highly technical language and be complex, which would likely result in low *comprehensibility*. Thus, the challenge in producing explanations for ML systems lies in appropriately balancing the tradeoff between *comprehensibility* and *completeness*. This balance will be heavily influenced by the user to whom the explanation is being provided and the context in which it must be provided. When an explanation of an ML system must provide a higher level of *comprehensibility*, terms like "intelligibility", "comprehensibility", and "understandability" are often used synonymously with "interpretability".[11,47,58] When an explanation of an ML system must provide a higher level of *completeness*, the term "transparency" is often used interchangeably with "interpretability".[47,58]

2.1.1.1 Levels and Types of Explanation

In recognition of the challenge of providing explanations for ML systems that appropriately balance *comprehensibility* and *completeness*, the literature has defined several "levels" or "types" of explanation for ML models. Some researchers refer to these as levels or types of interpretability,

[1]Gilpin et al.[52] originally used the term "interpretability", but I have chosen to use "comprehensibility" to avoid ambiguity with my previous discussion of the concept of interpretability

but as interpretability generally takes the form of an explanation, I will consistently use the term explanation to describe approaches to interpretability throughout this work.

Although the terminology differs, many researchers have chosen to distinguish types of explanations based on whether they describe model *processes* or *behavior*.[12,20,32,47,48,52,54,57,59] Explanations of *processes* focus on elucidating aspects of the model training algorithm, parameter settings, and/or internal representation/structure (i.e., the mathematical relationships between inputs and outputs).[47,48] These are referred to as "transparent", "white-box", or "descriptive" explanations, as they typically provide detailed descriptions of the internal operations of a model.[47,48,54,57,59] These explanations can provide insights into why an ML system may dysfunction, or fail to operate as intended, and are most useful in the context of debugging, monitoring, and improving systems.[20] Thus, these explanations tend to prioritize *completeness*, at the possible risk of lower *comprehensibility*. Explanations of *behavior* generally focus on clarifying how a model relates inputs to outputs,[32,54,59] and may involve showing the influence of each input, revealing characteristics of similarly classified instances, and/or changes in inputs that would result in a change in output.[55] These may be referred to as "black box explanations", "observations", "justifications", or "persuasive explanations" as they offer reasons for a model's outputs, but generally do not contain information regarding the internal operations of the model.[32,48,54,57,59] These explanations can provide insights into why a system may misfunction, or produce unintended or undesired effects, and are typically most useful in ensuring that a system meets various evaluation criteria such as unbiasedness, justifiability, etc.[20] Thus, these explanations tend to have higher *comprehensibility*, but lower *completeness*; however, it should be noted that in some cases, an explanation of model *processes* may be requested if an explanation of model *behavior* does not satisfy a user's information needs.[59]

Researchers have also distinguished between explanations provided at the *global* and *local/instance* levels of models, with these terms being used consistently throughout the literature.[12,15,26,43,51,56,60–62] According to Adadi and Berrada,[51] an explanation for a model at the *global* level "facilitates the understanding of the whole logic of a model and follows the entire reasoning leading to all the different possible outcomes". More generally, the goal of *global*-level explanations is to help users develop mental models of a model and how it works. These explanations can include information regarding the training information, the architecture and algorithms, the functional-level performance descriptions (e.g., accuracy), and boundary conditions and failure modes, i.e., information on what the model cannot do or does not perform well on.[61] These explanations tend to have high *completeness*, but it is generally challenging to improve the *comprehensibility* of these explanations. An explanation for a model at the *local* level provides the reasoning behind a specific model output or group of outputs.[51,62] This is also sometimes referred to as the *instance* level, which is the term adopted in this work. While *instance*-level explanations are also aimed at helping users gain mental models of the system, they focus on helping a user understand and interpret specific model outputs.[61] These explanations can achieve high levels of *comprehensibility*, but generally lack *completeness*. Both *global*- and *instance*-level explanations can be aimed at explaining either model *processes* or *behavior*, although typically *global*-level explanations describe model *processes* and *instance*-level explanations describe model *behavior*. Hall et al.[15] suggest that the best explanations for ML models will likely come from a combination of both *instance*- and *global*-level explanations.

2.1.2 Interpretability Approaches

As mentioned previously, I view approaches to interpretability as forms of explanation and thus consistently use the term explanation in reviewing the literature. However, it should be noted that outside of this work the term interpretability is used more frequently when describing and classifying various approaches. Although multiple prior attempts have used differing terminology to classify approaches to explanation,[15,20,26,28,47,51,52,62,63] it is generally agreed that there are at least two main categories of explanation approaches: *integrated* and *post-hoc*.[20,28,65,31,47,50,51,56,62–64] In the sub-sections below, I define each category and provide further sub-categories of approaches. I provide general descriptions of the various sub-categories and only give examples as is necessary to distinguish between the categories. For more examples of approaches, I recommend referring to one of the several literature reviews/surveys available.[26,28,51,52,63] It should be noted that many of the approaches in these categories make claims of interpretability that are not substantiated by empirical user studies.

2.1.2.1 Integrated Explanation Approaches

Following Došilović et al.,[28] I define *integrated* explanation approaches as those approaches that are transparency-based—that is, they are aimed at describing model *processes* and are generated as part of the learning/training process. These approaches generally provide *global*-level explanations. Combining insights from Došilović et al.[28] and Gilpin et al.,[52] I sub-divide these approaches into *pure transparent*, *hybrid*, and *explanation-producing* approaches.

In *pure transparent* approaches to explanation, the family of models that can be used is restricted to those that are considered transparent, or models whose internal mechanisms can be understood.[28,47] Typical examples of such model families include decision trees, Naïve Bayes

models, logistic regressions, and linear regressions, among others. The model itself can serve as an explanation and may be referred to as an "intrinsically interpretable model", "inherently interpretable model", "intelligible model", "comprehensible model", "transparent model", "transparent-box model", "white-box model", or "glass-box model".[26,29,51,56,66]

Pure transparent approaches will generate explanations that have high *completeness*, but the level of *comprehensibility* will depend on the complexity of the model. It is generally accepted that for *pure transparent* approaches, accuracy comes at the cost of *comprehensibility*.[51] For example, a linear regression model using hundreds of features or highly engineered features may exhibit high accuracy, but the complexity of the model leads to decreased *comprehensibility*. In *hybrid* approaches, transparent model families are paired with models whose internal mechanisms are generally considered to be opaque, i.e., "black-box" models, to produce models that sacrifice some *comprehensibility* to achieve better accuracy.[28] Again, the model itself typically serves as the explanation in these types of approaches. An example of a *hybrid* approach from Došilović et al.[28] combined logistic regression and support vector machine (SVM) approaches to credit scoring.

In *explanation-producing* approaches, models that are considered to be "black-boxes" are specifically built to provide explanations that improve the transparency of their internal mechanisms.[52] These approaches typically apply to neural networks, such as those that learn disentangled representations,[52] and the balance between *completeness* and *comprehensibility* will vary by approach.

2.1.2.2 Post-hoc Explanation Approaches

Post-hoc explanation approaches involve separating the tasks of model learning and explanation, i.e., applying explanation methods after the model learning/training process.[51,62] These approaches are sometimes referred to as "reverse engineering" approaches to explanation

as they involve a level of model reconstruction.[26,51] These explanations may be in the form of visualizations, natural language or text, rules, examples, and various other formats.[20,47,51] *Post-hoc* explanation approaches can be sub-divided into *model-specific* and *model-agnostic* approaches.[15,28,51,62,65]

Model-specific explanation approaches are only applicable to specific models as they rely on idiosyncrasies of the model's internal mechanisms.[28,51,62] These explanations can aim to describe model *processes* and/or *behavior* and can be provided for both the *instance*- and *global*-levels, although *global*-level explanations tend to be more common. Thus, the level of *comprehensibility* and *completeness* of a *model-specific post-hoc* explanation will vary by approach. It should be noted that all *integrated* explanation approaches mentioned in section 2.1.2.1 are also *model-specific*, but they are not *post-hoc*. An example of a *model-specific post-hoc* explanation approach is given by Barakat et al.,[25] who used model-specific techniques to extract rule-based explanations from an SVM classifier for predicting diabetes.

Model-agnostic explanation approaches are not tied to any specific model or algorithm, i.e., they treat the original model as a "black-box".[28,51,62,64] They generally operate by analyzing only the inputs and outputs of the original model and thus describe model *behavior*.[28,62,64] The *model-agnostic* approaches can be provided at both the *global*- and *instance*-levels, although *instance*-level explanations are more commonly seen. These approaches tend to produce explanations with lower *completeness* than other explanation approaches; however, they can typically provide high *comprehensibility* and offer the attractive advantage of being generalizable. More specifically, *model-agnostic post-hoc* explanation approaches provide general explanation formats that allow for customization to fit user information needs, enable comparisons of different models, and facilitate the process of switching out a model in a deployed ML system.[64] *Model-*

agnostic post-hoc explanation approaches can be loosely grouped by the technique used to generate the explanation: 1) visualizations (e.g., partial dependence plots, individual conditional expectation), 2) knowledge extraction (e.g., rule-extraction, model distillation), 3) feature influence methods (e.g., sensitivity analysis, feature importance/attribution), and 4) example-based (e.g., prototypes and criticisms, counterfactual explanations).[51]

2.1.3 Evaluating Interpretability Approaches

As what constitutes a good explanation for an ML model is both user- and context-dependent, it is unsurprising that the literature provides no standard approach to evaluating model explanations that claim to facilitate interpretability. Doshi-Velez and Kim[43] have posited that the evaluation of approaches to interpretability should match the claimed contribution. For example, if the aim of an explanation approach is to make a model useful in some context or application, then the explanations should be evaluated with respect to that application (e.g., explanations for a model to assist in medical diagnosis should be evaluated by having doctors use the system to make diagnoses). They proposed a simple 3-level taxonomy for approaches to interpretability evaluation: 1) *application-grounded*, 2) *human-grounded*, and 3) *functionally-grounded*. These general categorizations provide a useful framework for discussing general approaches to evaluating interpretability. Figure 3 provides an overview of each approach, which are discussed in detail in the next few paragraphs. In this section I discuss only general approaches to evaluating interpretability and I discuss specific studies evaluating interpretability within the healthcare domain in section 2.3.2.

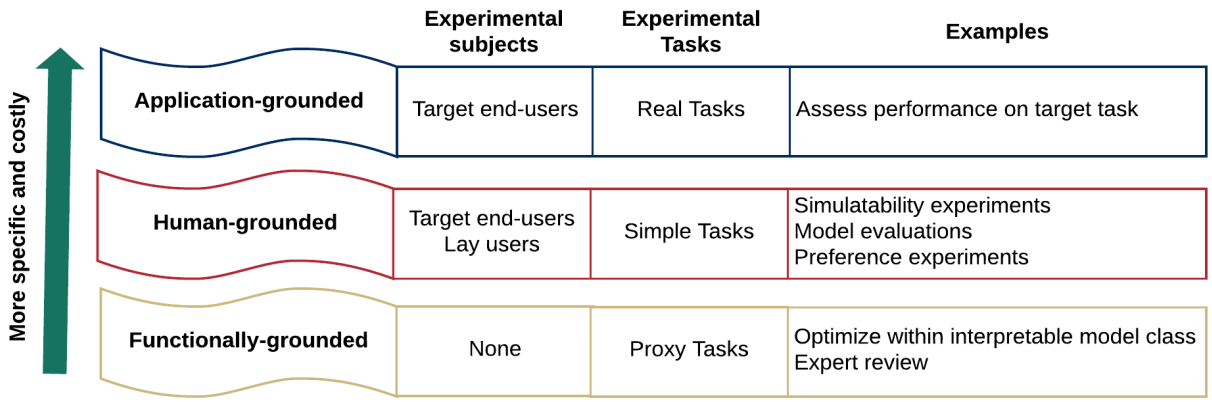


Figure 3. General approaches to interpretability evaluation (adapted from Doshi-Velez and Kim[43]). A summary of the three general approaches to interpretability evaluation with example experiments for each approach. Approaches are distinguished by the type of experimental tasks and subjects involved, and get more costly as the level of user-involvement and experimental complexity increases.

Application-grounded approaches involve experiments in which real humans, i.e., target end-users, use real applications to perform the intended end-task of the application. This allows for the evaluation of explanation quality within the context of its intended use. The suggested baseline for these types of evaluations is how well human-produced explanations assist in other humans completing the task (i.e., the gold standard is the success rate of completing the task using explanations provided by humans). These approaches are centered on the idea that if a system that employs an explanation approach has practical utility, then the explanations must satisfy the demand for interpretability. It should be noted that these are the most demanding evaluations to perform, requiring significant time, effort, and expense to complete.[43] Thus, few evaluations of interpretability employ an application-grounded approach, but some can be seen in the literature on explanations for recommender systems.[67]

Human-grounded approaches involve experiments in which real humans use explanations to perform simplified tasks that maintain the essence of the target application. These types of

experiments aim to test general notions of explanation quality and can usually be performed with lay users when experiments with target end-users prove logistically challenging (e.g., highly trained domain experts pose logistical challenges as they generally have smaller recruitment pools and higher compensation requirements).[43] As these approaches are typically simpler to employ than application-grounded approaches, they are the most commonly used in the interpretability evaluation literature when humans are involved. Simulatability experiments, i.e., does the explanation approach allow a human to easily predict a model's output for a given input, are quite popular in the literature.[20,32,56] The general motivation for this measure is that if an explanation approach has allowed a human to build a robust, accurate understanding of a model, then they should be able to simulate the model's behavior. Other common human-grounded approaches in the literature can be loosely categorized as model evaluation experiments (e.g., impact on ability to identify model errors and/or select best model), effectiveness experiments (e.g., impact on user ability to make decisions with the model), confidence/trust experiments (e.g., change in user prediction before and after seeing the model prediction), and preference experiments (e.g., user rates the quality of different explanation formats).[20,67,68] Efficiency experiments, i.e., measuring how long it takes a user to comprehend different explanations or perform tasks using explanations, are also human-grounded approaches, but are rarely seen in the literature.[26,67]

Functionally-grounded approaches involve no human experiments, and instead use some formal definition of interpretability as a proxy of explanation quality. An example of a possible proxy would be a family of models whose interpretability has been validated in human experiments (e.g., decision trees have been validated in some contexts).[43] Researchers can optimize the performance of an approach based on that proxy and use the proxy to substantiate their claims for interpretability (e.g., optimizing the performance of decision trees on some task could claim to be

an interpretable approach). Much of the literature on interpretability appears to take this approach to justify claims for interpretability, but it is debatable whether the current suggested proxies are appropriate. For example, many methods claim interpretability by adopting *pure transparent* or *hybrid* approaches to explanation and then using model size as a measure for explanation complexity.[26,43] These approaches suggest that the size and transparency of a model can serve as proxies of explanation quality. However, studies have found that end-user preferences for smaller or larger models is context-dependent, and the comprehensibility of various model families depends on the end-user (e.g., an end-user with no statistical background may not find a sparse linear regression comprehensible).[29,69] Thus, a small transparent model will not always be an appropriate proxy for explanation quality. Another proxy for explanation quality sometimes seen in the literature is agreement with knowledge about the underlying model and/or the domain problem, i.e., if the explanation seems reasonable according to modeling and/or domain experts.[52] This typically involves a few experts reviewing explanations, but is a far more informal and small-scale evaluation than a human-grounded approach and is perhaps one of the weakest approaches for evaluating interpretability. The general challenge in functionally-grounded approaches lies in identifying appropriate proxies for explanation quality, particularly because there are limited studies on user-based measures of interpretability and relevant concepts such as comprehensibility are difficult to quantify.[11,28,43,68]

2.2 User-centered Explanation Design and Evaluation for AI Systems

This section provides an overview of the relevant literature on user-centered explanation design and evaluation for AI systems. This section is not meant to serve as a comprehensive review

of the literature, and instead focuses on summarizing some existing design and evaluation frameworks and their limitations. A summary of guidance on user-centered explanation design and evaluation that is relevant to the proposed framework is also provided.

A framework proposed by Wang et al.[40] relies on theories of human reasoning and explanation, and highlights specific elements of AI explanation that support these processes and mitigate errors. The framework is shown graphically with a brief description in Figure 4. The framework promotes explanation design by linking specific AI explanation techniques and elements to the human cognitive processes and patterns they can support (e.g., "what if" type explanations support counterfactual reasoning; information about the prior probability can help mitigate confirmation bias). In a follow up position paper,[41] the researchers used the framework to theoretically justify the use of specific explanation types to support various user goals based on how users may generally reason about the goal. For example, a user trying to identify a specific cause for a particular system outcome might employ contrastive reasoning, which can be supported by "why not" type explanations that provide information about why an alternative system outcome was not produced. Although some examples are provided, the authors provide limited guidance on how to connect reasoning processes to specific AI elements and techniques. The framework also does not consider the specific type of user when considering goals and cognitive processes and it does not account for the environment in which an explanation is being provided. Moreover, the framework links reasoning processes to a non-comprehensive list of AI explanation elements and techniques that *currently* exist, which provides limited guidance on how user reasoning can inform the design of new displays for existing explanation algorithms as well as for new explanation algorithms.

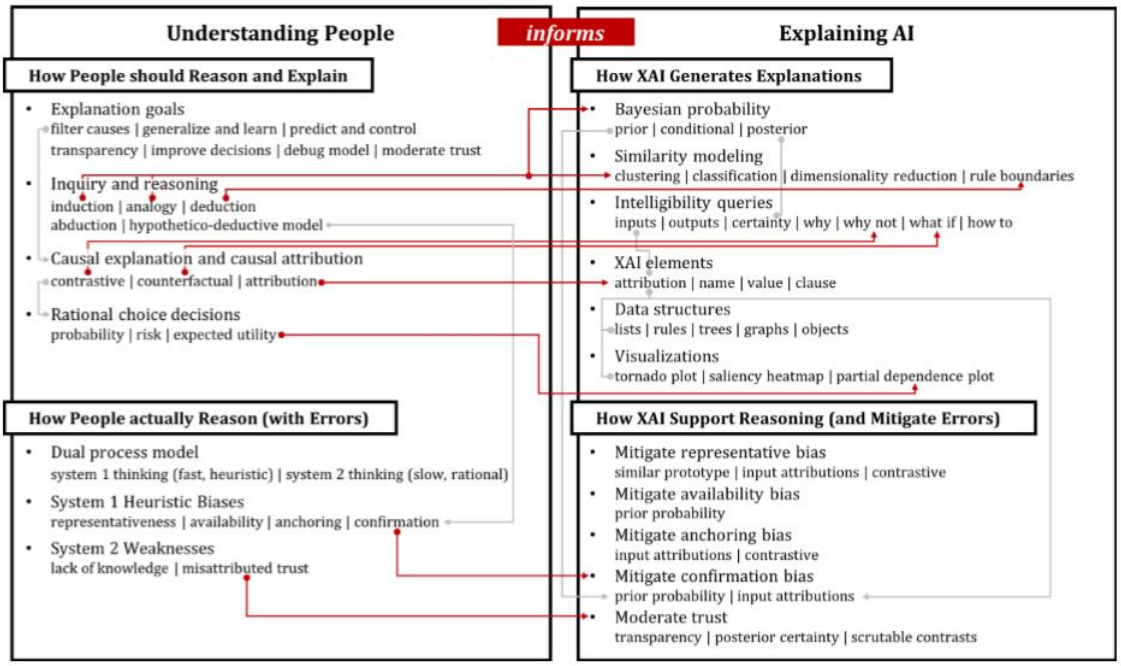


Figure 4. Conceptual framework for reasoned explanations that describes how human reasoning processes (left) inform explainable AI techniques (right) from Wang et al.[40] "Points describe different theories of reasoning, explainable AI techniques, and strategies for designing explainable AI. Arrows indicate pathway connections: red arrows for how theories of human reasoning inform explainable AI features, and grey for inter-relations between different reasoning processes and associations between explainable AI features. Only some example pathways are shown. For example, hypothetico-deductive reasoning can be interfered by System 1 thinking and cause confirmation bias (grey arrow). Confirmation bias can be mitigated (follow the red line) by presenting information about the prior probability or input attributions. Next, we can see that input attributions can be implemented as lists and visualized using tornado plots (follow the grey line)." (Image and caption taken directly from Wang et al.[40])

A framework proposed by Ribera and Lapedriza[42] is based on theories that describe explanation as a social interaction. The framework is shown graphically with a brief description in Figure 5. Their framework focuses on understanding the explainee (i.e., the user) needs and providing explanations that both meet those needs and follow Grice's maxims of conversation[70] (quantity, quality, relation, manner—in short, only be as informative as needed, be truthful, be

relevant, and be perspicuous). The framework describes three general user types based on their background and relationship to the AI system: 1) AI experts—researchers who develop an AI system, 2) domain experts—specialists in the area in which the system is being used (e.g., physicians, lawyers, etc.), and 3) lay users—the recipients of the final decisions of the system (e.g., a patient that has been diagnosed). They combine these user types with Grice's maxims to identify specific explanation goals (why), the content to include in an explanation (what), the type of explanation or explanation approach (how), and suitable evaluation approaches for each user type. Although the proposed framework helps elucidate general explanation design ideas to support the goals for each user type, it includes only a select set of the available concepts on explanation design available in the model interpretability literature and it does not consider the environment in which the explanation is being provided to a user. Additionally, the framework is difficult to utilize when the user types overlap (e.g., a lay-user who is also a domain expert).

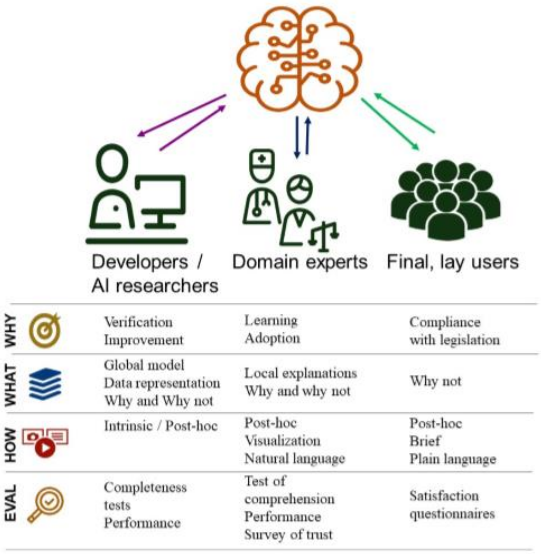


Figure 5. User-centric framework based on Grice's conversation maxims from Ribera and Lapedriza.[42] "The system targets explanations to different types of user, taking into account their different goals, and providing relevant (Grice's 3rd maxim) and customized information to them (Grice's 2nd and 4th maxim). Evaluation methods are tailored to each explanation." (Image and caption taken directly from Ribera and Lapedriza[42])

Mohseni et al.[44] reviewed existing literature on explainable AI from a variety of domains to develop a general framework for the design and evaluation of explainable AI systems that considers the type of users and their primary goals and needs. The authors identify three user types similar to the three types proposed by Ribera and Lapedriza[42]: 1) AI novices, or end-users of AI products that have limited knowledge of ML, 2) data experts, or data scientists and domain experts who use ML approaches but generally lack in-depth expertise, and 3) ML experts, who design and have a strong theoretical understanding of ML algorithms. The authors suggest designing explanations by identifying the intended user of an explainable AI application, choosing an AI application that meets the targeted user's primary goals and needs, choosing the explanation type and format that supports the user type and intended application, and finally performing user-evaluations of the explanations. Mohseni et al.[44] expand upon this suggested approach by proposing a three-level nested model to design and evaluate an explainable AI system, where each level builds upon the work of previous levels. Figure 6 depicts and briefly describes the model.

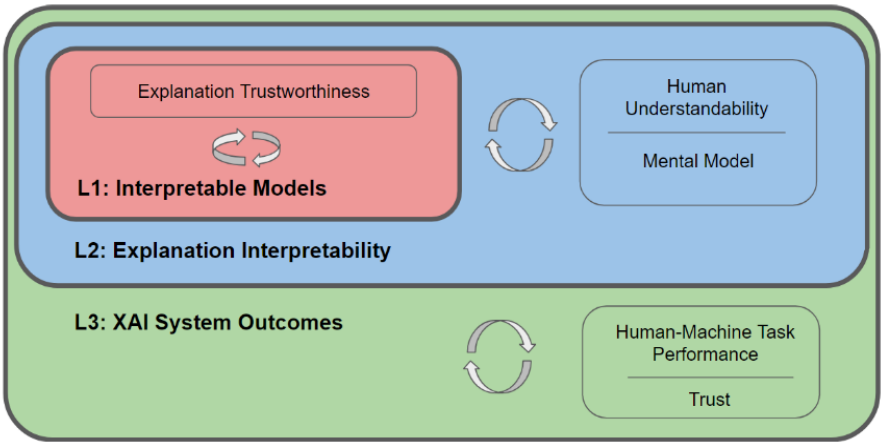


Figure 6. Three-level nested model to designing and evaluating an explainable AI system from Mohseni et al.[44]
"The innermost layer (Red) presents design and evaluation of interpretable ML algorithms. The middle layer (Blue) shows design and evaluates human understandable explanations and explainable intelligent interfaces and agents. The outer layer (Green) demonstrates evaluation of explainable AI system outcomes with end-users." (Image and caption taken directly from Mohseni et al.[44])

The goal at the lowest level (Interpretable Models Level) is to design understandable models, which usually involves ML experts who want to evaluate the trustworthiness and reliability of an explanation method, usually utilizing computational measures such as comparison to an interpretable model. The goal of the middle level (Explanation Interpretability Level) is to design understandable explanations that satisfy target user usability needs, which usually involves subjective evaluations of target user satisfaction with and understanding of the system. The goal at the highest level (Explainable AI System Outcomes Level) is to evaluate the ability of the explainable AI system to satisfy target-user needs, which usually involves domain-specific subjective and objective measures of the system impact on user task performance and perceptions of the system. These three levels and proposed evaluation metrics closely align with the functionally-grounded, human-grounded, and application-grounded approaches to evaluating interpretability proposed by Doshi-Velez and Kim,[43] respectively. While this proposed framework is useful when discussing the big picture of user-centered design and evaluation of explainable AI systems, it provides limited guidance on explanation designs that could support user needs at each of the three levels.

In addition to the previously mentioned frameworks, a few other authors have provided useful insights that can guide user-centered explanation design and evaluation. Ras et al.[20] offer a categorization of users of ML systems that further expands upon the general user types proposed by Ribera and Lapedriza[42] and Mohseni et al.[44] The authors define two broad categories of users based on expertise: 1) expert users who are responsible for implementing an ML system and who typically have some knowledge about the inner workings of an ML system, and 2) lay users who are the people for which an ML system is built and who are not expected to have knowledge about the inner workings of an ML system. The two categories are further sub-divided into specific types

of users, which loosely represent the various relationships a user may have with an AI system. For each categorization, Ras et al.[20] identify possible goals and concerns that may prompt the user to ask for explanations. These goals and concerns fall under Samek et al.'s[71] four broad categories of reasons why users seek explanations of AI systems: 1) verification, 2) improvement, 3) learning, and 4) compliance. Verification includes examining how decisions/suggestions are made by the system to ensure it is operating as expected. Improvement can be closely tied to verification and covers activities related to improving the system performance and efficiency. Learning refers to any activity where the user seeks to extract knowledge from the system. Compliance is also closely tied to verification and relates to any activities aimed at ensuring the system adheres to an established legal, moral, or other societal standard. Table 2 combines the insights from Ras et al.[20] and Samek et al.[71] to provide definitions and possible explanation goals for different user categories, which are not intended to be mutually exclusive (i.e., a user may belong to more than one category).

Other insights that can guide user-centered explanation design and evaluation come from the previously discussed work on interpretability evaluation by Doshi-Velez and Kim.[43] The authors hypothesize several factors that may influence user explanation needs, highlighting the importance of considering user expertise and environmental factors (e.g., time constraints) when completing a task. Additionally, the authors define cognitive chunks as the basic units of explanation, and suggest that the form, number, level of compositionality (i.e., how chunks are organized), and relationship (e.g., combination of chunks in linear or nonlinear way) of these chunks may differ based on user explanations. These concepts demonstrate more general design considerations than those introduced in the framework by Wang et al.,[40] yet more specific than those suggested by Ribera and Lapedriza.[42]

Table 2. Categories of users and explanation goals (adapted from Ras et al.[20] and Samek et al.[71])

Main Category	Sub-category	Description and Concerns	Explanation Goals
Expert user	Engineer	• Have detailed knowledge about mathematical theories and principles behind a system • Concerned with developing and improving ML algorithms/models	• Verification E.g., debugging models • Improvement E.g., identifying ways to improve existing models
	Developer	• Focus on building ML systems for lay people • Often utilize off-the-shelf ML algorithms • Concerned with satisfying various use cases of the ML system	• Verification E.g., model behavior alignment with use case criteria • Improvement E.g., hyperparameter tuning
Lay user	Owners	• Acquire ML system for use • Individuals and organizations • Concerned with evaluating capabilities of system	• Verification E.g., justification of predictions, malfunction rate • Compliance E.g., liability/safety concerns
	End-users	• Individuals expected to use the ML system as part of personal and/or professional activities • Concerned with understanding capabilities of system	• Verification E.g., justification of prediction, reliability concerns • Learning E.g., actionable outcomes, assistance in completing task
	Data subjects	• Individuals or entities whose information is being processed or who are otherwise directly affected by the ML system • Concerned with system impact on self	• Compliance E.g., adherence to ethical principles, privacy concerns
	Stakeholders	• Other individuals or organizations who claim an interest in the ML system, but are not directly connected to its development, use, or outcome • Concerned with system impact in general	• Compliance E.g., adherence to ethical principles, liability concerns

2.3 Interpretability in Healthcare

2.3.1 Motivations for Interpretability

The high stakes and complex nature of healthcare motivates ML-based applications which assist healthcare providers in achieving various goals.[12] Thus, much of the literature motivates the need for model interpretability in healthcare by claiming that it is integral to the usability,

acceptability, and trustworthiness of an ML model.[6–8,10–12,29] Although these motivations are somewhat vague, most are linked to real-world goals in which criteria such as informativeness, accountability, liability, justifiability, etc. must be satisfied.

The most common goal for a provider utilizing an ML model is to provide better and more effective care for a patient.[12] In most cases, ML models are proposed as a tools to help providers analyze patient data and derive insights that can guide clinical decision-making.[4,12,29,34] Effective use of an ML model in practice requires a healthcare provider to assimilate knowledge from the model and reconcile it with prior knowledge and missing contextual information to make informed care decisions. As clinical decisions made with the assistance of an ML model may affect the lives of patients, it is essential that providers be able to validate the information provided by the model.[4,8,34,51] Moreover, healthcare practitioners are legally and ethically responsible for any care decisions made based on ML model information, and will therefore be unlikely to adopt or deploy ML models that cannot justify their outputs or be vetted for potentially critical errors or data bias.[6,8,11,12,21,29,34] Thus, when ML models are used to assist in providing patient care, interpretability may be demanded to allow providers to derive actionable insights from the model, verify model outputs before acting on them, and defend care decisions based on the ML model.

The demand for model interpretability may also present itself even when ML models are not used directly in clinical practice. For example, it is generally accepted that ML models can be improved by integrating knowledge and feedback from domain experts into the learning/development process.[6,7,72] Model interpretability approaches can serve as tools that facilitate conversations between healthcare providers and model developers. These conversations could lead to improved models for use in healthcare. Alternatively, ML models may be used in healthcare as data-driven approaches to generating new knowledge that could help advance the

field.[7,12,34] By uncovering correlations between patient characteristics and outcomes of interest, ML approaches to predictive modeling can assist domain experts in causal reasoning and hypothesis generation.[34,69,73] In this case, the explanation of a model should assist domain experts in identifying new predictively accurate explanatory variables to study,[7,73] potentially leading to new therapies and interventions that lead to improved outcomes and lower costs.[34]

2.3.2 Explanation Approaches and Evaluations

There has been a recent surge in publications of high-performing models for healthcare applications that also make a claim of interpretability (Figure 7). Table 3 uses the terms introduced in section 2.1 to summarize some of the interpretability approaches seen in this recent set of work. This table only captures a subset of the relevant literature, but it does offer insight as to model interpretability research from the ML community that has been used in healthcare applications.

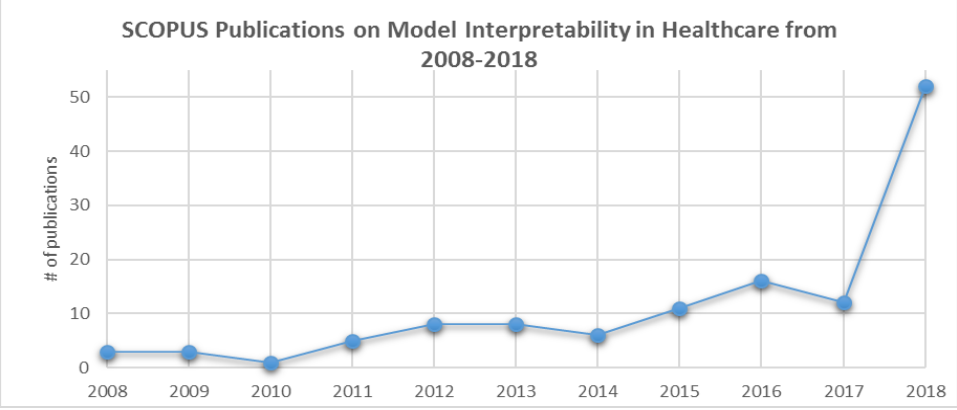


Figure 7. SCOPUS publications on model interpretability in healthcare from 2008-2018. Aggregate numbers were generated using the search query ((TITLE-ABS-KEY(("predictive model" OR "artificial intelligence" OR "machine learning") AND ("healthcare" OR "medicine") AND ("transparent" OR "intelligible" OR "explainable" OR "explanation" OR "interpretable" OR "comprehensible" OR "understandable"))). The query was run on November 29, 2018.

Table 3 shows that researchers have recently begun to explore alternative approaches to interpretability other than the use of logistic regression and other comprehensible models. However, human evaluations with end-users are rarely performed. Such studies are particularly important in healthcare, where providers are already overwhelmed by vast amounts of data. It is vital that ML models and explanations be delivered in a manner that does not exacerbate this problem.[6,12] Additionally, the intended user should be satisfied with the information provided.[6,12] Based on the literature survey, no human-evaluation studies of model explanations in healthcare have fully addressed these issues. Krause et al.[74] performed a human evaluation of a custom visual explanation approach, but the target end-users of their system were data scientists/analysts and not healthcare providers. Lundberg et al.[75] performed a small-scale study to evaluate whether explanations for predictions improved anesthesiologists' ability to predict hypoxemia risk during surgery, but did not assess provider perceptions of satisfaction with the system and explanations.

Table 3. Model explanation approaches and evaluations in the recent healthcare literature

			Katuwal & Chen[10]	Yang et al.[16]	Lundberg et al.[75]	Caruana et al.[76]	Choi et al.[77]	Che et al.[78]	Barakat et al.[25]	Luo et al.[79]	Jovanovic et al.[14]	VanBelle et al.[80]	Soininen et al.[81]	Krause et al.[74]	Letham et al.[17]	Kunapuli et al.[82]	Yang et al.[83]	Sha and Wang[84]	Valdes et al.[85]	Liu et al.[86]
Explanation Approaches Used	**Level**	**Global**		X					X		X	X		X	X	X	X		X	
		Instance	X	X	X	X	X	X		X			X	X			X	X		X
	Integrated	**Pure transparent**		X							X	X	X			X			X	
		Hybrid													X					
		Explanation-producing					X											X		
	Post-hoc	**Model-specific**				X			X								X			
		Model-agnostic	X	X	X			X		X				X						X
Evaluation	**Functionally-grounded**		X	X		X	X	X	X	X	X	X	X		X	X	X	X	X	X
	Human-grounded																			
	Application-grounded				X									X						

3.0 Proposed Framework for Designing User-Centered Explanations

In this chapter, I present my proposed framework for user-centered explanation design for ML-based systems in healthcare, which is depicted in Figure 8. The framework was inspired by the frameworks of Wang et al.[40] and Ribera and Lapredizа[42] and incorporates other guidance on user-centered explanation design for AI systems. The purpose of this framework is to propose a general approach to user-centered explanation design that can be applied to the adoption of existing explanation approaches and to the development of new approaches. Therefore, specific design suggestions (e.g., a specific explanation approach or presentation method) are not included and the examples provided are not meant to be comprehensive. However, the examples provided in the framework encompass many of the ideas that appear in the literature on interpretability and user-centered explanation design and evaluation.

Prior to presenting the framework, it is important to clarify its scope and limitations. The proposed framework was developed to design explanations for empirically-based predictive models, or data-driven models based on statistical associations that aim to minimize prediction error.[73] It is not intended to be used for explanatory models, or theory-driven models that aim to test causal relationships between variables and that may be used in prediction tasks. A more in depth discussion on the differences between predictive and explanatory modeling is provided by Shmueli.[73] Additionally, the proposed framework does not explicitly consider how the use of specific data types or models may influence explanation design and interpretation. For example, models that use image data would have a very different space of possible explanation designs than models that use text data. Similarly, the space of possible explanation designs would change based on the specific model used, as different models will have different model-specific approaches to

explanation (e.g., Gini importance to show feature influences in a Random Forest model). Moreover, users with knowledge of modeling approaches and their limitations may interpret predictions and explanations of specific models differently. Thus, it is important to acknowledge the potential role of specific data types and models in explanation design and interpretation; however, these considerations were outside the scope of the proposed framework.

The framework is described in detail in Section 3.1. Section 3.2 provides guidance on how the framework can be applied within the larger context of the design and evaluation of explainable AI systems.

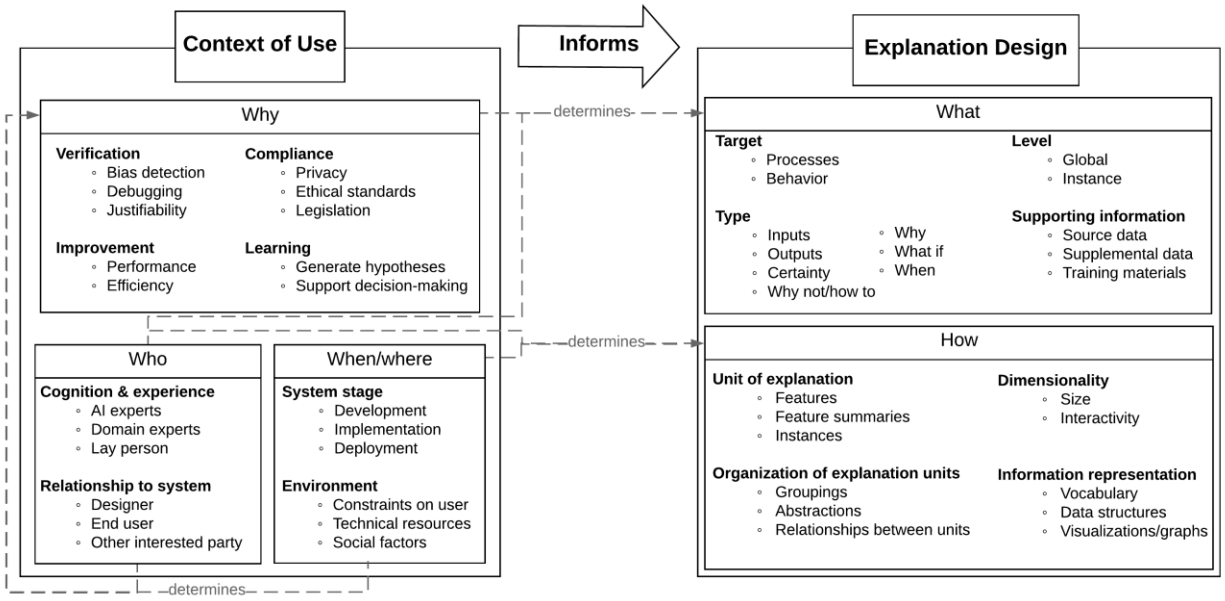


Figure 8. Proposed framework for designing user-centered explanations. The framework was inspired by the frameworks of Wang et al.[40] and Ribera and Lapredizа[42] and incorporates insights from work on explanations by Ras et al.[20], Samek et al.,[71] Lim et al.,[41] and Doshi-Velez and Kim.[43]

3.1 Description of Framework

The framework suggests that user-centered explanation design should be informed by the entire context of use of an explanation—that is, explanation design should consider not only *who* an explanation is being provided to and *why* they want that explanation, but also *when* and *where* that explanation will be used. Answering *who, why, when,* and *where* about the use of an explanation can be used to inform *what* information the explanation needs to contain (i.e., the content) and *how* that information needs to be provided (i.e. the presentation). As indicated by the relationships between the target questions indicated by grey dashed lines in Figure 8, these target questions are not orthogonal and are often co-dependent in that the answers to one question can and will be determined by the answers to other target questions. This introduces a partial ordering to the way in which target questions should be answered. More specifically, *who* an explanation is provided to and *when/where* that explanation is being provided should be answered first, as this information can then generally be used to answer *why* the explanation is needed. Similarly, *who* an explanation is provided to and *why* the explanation is needed typically determines the answer to *what* needs to be in the explanation. Finally, *how* the information in the explanation is presented to a user can generally be answered by *who* the explanation is provided to and *when/where* the explanation is being provided. As shown in Figure 8, each target question (*who, why, when, where, what,* and *how*) is associated with general factors that should be considered for each question (e.g., cognition and experience for *who*) along with some specific examples for each factor (e.g., AI expert). These are further discussed in the paragraphs below, following the suggested ordering for answering the questions.

The answer to the target question *who* plays a major role in answering several other target questions and should be answered first. Prior work has tried to create categories of users to define

explanation design needs, but as Ras et al.[20] noted, users often don't fall into a single category. I assert that users can generally be defined by two aspects: 1) user cognition and experience (e.g., knowledge, capabilities, influence of prior experiences, etc.) and 2) the user's relationship to the system at the time the explanation is being provided. Ribera and Lapedriza's[42] classifications of AI experts, domain experts, and lay persons capture the main categories of user cognition that appear in the literature. Ras et al.'s[20] sub-categorizations of users (engineer, developer, owner, end-user, data subject, stakeholder) capture the various relationships a user may have with an AI system. In Figure 8, these sub-categorizations are generalized into the role of designer (engineer, developer), end user, and other interested party (owner, data subject, stakeholder). Defining users using these two dimensions overcomes the problem of trying to create mutually exclusive user categories to define needs. A user may have several different relationships with the system over time, and thus their explanation needs may change with varying roles.

The *when* and *where* target questions are closely tied with the *who* target question, and also play a role in answering several other target questions. Perhaps the broadest classification of *when/where* an explanation is being used is related to the stage of the system, which often defines a user's relationship to the system (e.g., during development the user relationship to the system is often that of designer). Explanations required during system development, implementation, and deployment will likely differ in design due to the different environmental settings associated with each stage. More specifically, *when/where* can be answered by considering the environment in which the explanation will be used and how the explanation needs to be designed in order to support use within that environment. Specifically, environment will dictate the constraints on the user (e.g., available time and cognitive capacity), the available technical resources, and the user's perception of the system, which are all factors that may influence explanation design.

The *why* target question can often be answered by the answers to the *who* and *when/where* target question, as the user, their relationship to the system, and the environment in which they will be operating often affects why an explanation is sought. Although several prior works have identified various user needs and goals that drive the need for explanations, most of these can be captured in Samek et al.'s[71] four broad categories of reasons *why* explanations of intelligent systems are required: 1) verification, 2) improvement, 3) learning, and 4) compliance. Verification includes examining how decisions/suggestions are made by the system to ensure it is operating as expected, which may include activities such as detecting biases, finding and debugging errors, and ensuring that system reasoning aligns with domain knowledge (justifiability). Improvement covers activities related to improving the system performance and efficiency, which may include things such as incorporating domain knowledge to reduce biases in or improve generalization of the system, comparing and selecting between models, and improving system response times. Learning refers to any activity where the user seeks to extract knowledge from the system, which may include identifying previously unknown data patterns, generating/testing new hypotheses, and improving decision-making accuracy or speed. Finally, compliance relates to any activities aimed at ensuring the system adheres to an established legal, moral, or other societal standard. It should be noted that these are not mutually exclusive categories (e.g., explanations for verification are also often used to guide improvement activities). When users request explanations in the context of decision-making, they are generally requesting explanations for verification (e.g. support for a specific decision suggested by the system) and/or explanations for learning (e.g., knowledge to support a decision-making process).

The *what* target question refers to the content that needs to be included in an explanation. This can generally be determined by the answers to the *who* and *why* target questions, but

additional context and inquiry with the target users may be required. Depending on *who* is receiving the explanation and *why* they require it, the explanation design may need to be targeted at explaining either the internal processes of a system (i.e., how it specifically relates inputs to outputs) or its general behavior (i.e. input/output relationships only) and the explanation may need to be provided at the global (i.e., explains the entire model or system) or instance (i.e., explains a single prediction) level. The target and level of the explanation design can generally be determined by the type of explanation the user is seeking. Lim et al.[41] provide a useful taxonomy of explanation types based on the intelligibility query they aim to answer: 1) "input" explanations, which provide information on the input values being used by a system; 2) "output" explanations, which provide information on specific outcomes/inferences/predictions; 3) "certainty" explanations, which provide information on the uncertainty of a certain output; 4) "why" explanations, which provide information on how a system obtained an output value based on certain input values (i.e., model traces or complete causal chains); 5) "why not"/"how to" explanations, which provide information on why an expected output was not produced based on certain input values (i.e., contrastive explanations, counterfactuals); 6) "what if" explanations, which provide information on expected changes in outputs based on certain changes in the input (i.e., explanations that permit outcome simulations); and 7) "when" explanations, which provide information on which circumstances produce a certain output (i.e., prototype or case-based explanations). It should be noted that these categories are not mutually exclusive and can be combined in various ways (e.g., it is possible to provide an "input"/"output"/"certainty"/"why not" explanation). Depending on user cognition and needs, the explanation may also need to be supported by additional information such as source data (e.g. raw data the model was built from), supplemental data (e.g., data not included in the

modeling process but relevant to the situation or context), and training materials (e.g., information on model development or explanation interpretation).[40]

The *how* target question refers to the way in which the content of an explanation is presented to a user, which can generally be determined by the answers to the *who* and *when/where* target questions. Summarizing and expanding upon the work of Doshi-Velez and Kim,[43] the presentation of an explanation can generally be summarized using 4 main categories: 1) the unit of the explanation, or the form of the cognitive chunk being processed (e.g., raw features, feature summaries, images, or instances); 2) the organization of the explanation units, or the compositionality and relationship between the units, which may include groupings, hierarchical or relational organizations, or summary abstractions (e.g., free text summary of a combination of units); 3) the dimensionality, or processing size/levels of explanation information, which may include the overall size of an explanation and/or interactive exploration options; and 4) the manner in which information is represented, which includes the vocabulary, data structures, and visualizations used to express information. The specific choices in each of these four main categories will be determined by the user for whom an explanation is being provided (i.e., the *who*) and the environment in which it is being provided (i.e., the *when/where*).

3.2 Guidance on Application

It is useful to consider the application of the framework in the context of Mohseni et al.'s[44] three-level nested model to designing and evaluating an explainable AI system (see Figure 6 in section 2.2) and the taxonomy of evaluation approaches proposed by Doshi-Velez and Kim[43] (see Figure 3 in section 2.1.3). Specifically, the context of use portion of the framework can provide

valuable information for design and evaluation at all three levels. The context of use can be elicited using a variety of approaches involving target users, including but not limited to, interviews, workshops, surveys, site visits, focus groups, and/or contextual inquiry. Literature insights on target users and/or their environment can also help elucidate certain aspects of a context of use, but it is best to include some level of target user input when defining an entire context of use to inform design.

At the lowest level in Mohseni et al.'s[44] model (Figure 6, Interpretable Models Level, red layer), where experts develop new approaches to model interpretability and explanations are typically evaluated using functionally-grounded approaches (i.e., no human involvement), the framework can help developers consider the users and environments for which their approach might be best suited. This could assist developers in marketing their approach to the right audience (e.g., model developers, lay users) or to inform design requirements for the approach if the developers intended to target specific end-users or environments. For example, if developing an explanation approach that is intended to be used by lay persons who have limited knowledge of modeling processes, developers might want to ensure their approach focuses on explaining system behavior over system processes. The framework can also be helpful in considering which metrics to use when evaluating an approach. For example, if the approach is intended to provide real-time explanations in a dynamic environment (e.g., explanations for a CDSS), evaluating the computational efficiency of the approach would be vital.

The framework has direct applicability at the middle level in Mohseni et al.'s[44] model (Figure 6, Explanation Interpretability Level, blue layer), where the goal is to design explanations that satisfy target user usability needs and where human-grounded evaluation approaches are typically utilized. Specifically, using the framework to define a context of use helps elucidate the

explanation design requirements that need to be met. With a defined context of use, design requirements can be defined and existing explanation approaches can be assessed to determine whether they meet or can be adapted to meet the specified requirements. Preliminary explanation designs can then be proposed and refined in a series of human-grounded evaluations utilizing target users. The context of use and defined design requirements can help in selecting evaluation metrics to use in these studies. For example, if an explanation is intended to be used in a fast-paced environment, then an evaluation study might compare proposed explanation designs by the ease and speed with which target users can process an explanation.

At the highest level in Mohseni et al.'s[44] model (Figure 6, Explainable AI Systems Outcome level, green layer), the explanation design has been finalized, and human-grounded and/or application-grounded evaluation approaches are employed to evaluate whether the system satisfies target-user needs. These evaluations usually involve domain-specific subjective and objective measures of the system impact on user task performance and perceptions of the system. By using the context of use of a system to define target user needs, the framework can assist in designing evaluation studies. For example, consider a system that is intended to be used for decision-making in a high stakes environment (e.g., medical decision making). Users may require that the accuracy and effectiveness of such a system be thoroughly validated before they would accept or use the system. This would suggest that a series of human-grounded evaluations with target users evaluating the accuracy and effectiveness of the system would be useful evidence to support an application-grounded evaluation of the impact of the system.

In chapters 4-6, I demonstrate an application of the defined framework to design explanations for a model that predicts in-hospital mortality for pediatric ICU patients. In chapters 4 and 5, I demonstrate an application of the framework at the middle level of Mohseni et

al.'s[44] model by 1) defining a context of use for the model based on literature insights and past experiences, 2) suggesting preliminary explanation designs, and 3) refining the context of use and explanation designs utilizing human-grounded evaluation approaches. In chapter 6, I demonstrate an application of the framework at the highest level of Mohseni et al.'s[44] model by evaluating the predictive model with the refined explanation design to determine if the system satisfies target user needs.

4.0 Application of Framework to Suggest Explanation Designs for a Pediatric ICU In-hospital Mortality Risk Model

In this chapter I apply the proposed framework to the problem of predicting mortality risk for patients admitted to the pediatric ICU. Mortality risk prediction is a common application of ML in medicine.[87,88] In critical care, mortality prediction models have been used to establish performance benchmarks for outcome comparison and quality improvement initiatives, to define endpoints or illness severity adjustments in research studies, and to assist in clinical decision-making by providing early warnings of clinical deterioration.[89] With regard to using these models for clinical decision-making in the critical care environment, there has been growing interest in utilizing data mining techniques and ML approaches to build customized prediction models using data from local electronic health record (EHR) data repositories.[90–92] These models can be integrated into EHRs to provide real-time, individualized patient mortality risk predictions, which can assist critical care providers in surveilling patients for changes in acuity,[93] determining clinical priorities,[94] and providing support for making decisions about prognosis and treatment.[95,96] As predictive models are often based on incomplete information about a patient, use of a predictive model in decision-making challenges providers to integrate information from the model with their own clinical knowledge.[45,97] However, this process may require significant cognitive demands when providers do not understand the clinical basis for a prediction.[45] Thus, healthcare providers are unlikely to use a prediction model in decision-making without explanations.[45,93,97] Therefore, I applied the proposed framework to gain a better understanding of the potential benefit of providing user-centered explanations for mortality risk prediction models, specifically in the context of using the model to aid in decision-making processes. I focused the work on in-hospital

mortality risk for pediatric ICU patients, as there was a pre-existing dataset that could be used for model development.

In section 4.1, I describe the development and evaluation of an in-hospital mortality risk prediction model for pediatric ICU patients. In section 4.2, I apply the framework by utilizing insights from the literature and my prior experiences in developing predictive models to define a context of use for the model and identify promising explanation design requirements. In section 4.3, I present preliminary explanation designs for the model that will be refined in human-grounded evaluations with target users.

4.1 Development and Evaluation of the Mortality Risk Prediction Model

Data mining and ML approaches were utilized to develop a customized mortality risk prediction model for pediatric ICU patients at a single institution. As the main purpose of this work was to explore the utility of user-centered explanations for the model, a small, readily available dataset was utilized and no attempt was made to learn a best performing model. Section 4.1.1 describes the dataset and model development process while section 4.1.2 provides the results and final selected model.

4.1.1 Materials and Methods

4.1.1.1 Dataset Description

This work utilized a pre-existing dataset including all discharged patients with a pediatric ICU admission at the Children's Hospital of Pittsburgh (CHP) between January 1, 2015 and

December 31, 2016. Each hospitalization was treated as a separate encounter. The Institutional Review Board (IRB) of the University of Pittsburgh approved the use of the data for this dissertation work (PRO17030743).

For each encounter, the dataset included demographic information (age, sex, race), hospitalization data (time of admission and discharge), outcome data (discharge disposition and deceased date), assigned diagnoses, recorded locations, mechanical ventilation information, physical assessment measurements (vital signs, pupil reaction results, and Glasgow Coma Scale[98] (GCS) measurement), and laboratory test results. Encounters with a length of stay of less than 24 hours or unknown age, sex, or admitting diagnosis were excluded from the analysis. Only encounters with at least one recorded physical assessment measurement or laboratory test result were included in the dataset. The target outcome to predict was in-hospital mortality, which was defined as an encounter with a recorded deceased date that occurred on or prior to the recorded discharge date. The aim was to predict in-hospital mortality 24 hours prior to the event, and all data collected prior to the time of the prediction was utilized. For death cases, this included all data collected up to 24 hours prior to death and for control cases, this included all data collected prior to discharge.

4.1.1.2 Data Cleaning

The data cleaning processing was divided by categorical and numerical data types. Categorical data included sex, race, diagnoses, recorded locations, mechanical ventilation information, and pupil reaction results. Each categorical variable was mapped to a defined set of standard values. For sex, values were standardized to either “male” or “female”. For race, values were standardized to the six race/ethnicity categories defined by the Office of Management and Budget—“American Indian or Alaska Native”, “Asian”, “Black or African American”, “Native

Hawaiian or other Pacific Islander", "White", "Hispanic or Latino")[99]—with the addition of categories for "unknown" and "multiple" races. Diagnoses were recorded in *International Classification of Diseases* (ICD) Versions 9 and 10, and ICD-9 codes were mapped to ICD-10 whenever possible using the General Equivalency Mapping files available on the Center for Medicare and Medicaid Services website.[100] Locations were standardized to one of seven generic unit types: "Direct Admit", "Emergency Department", "pediatric ICU", "Other ICU", "Inpatient", "Outpatient", or "Operating Room". Left and right pupil reaction results were paired by timestamp and standardized to one of six possible values summarizing the results: "normal", "one sluggish", "both sluggish", "one nonreactive" "one sluggish, one nonreactive", "both nonreactive". Pupil reaction results that could not be paired (e.g., did not have both a left and right pupil reading with the same timestamp) were removed. After standardizing the possible values for each categorical variable, all duplicate results were removed.

Numerical data included age, length of stay, vital signs, GCS measurements, and laboratory test results. Laboratory test and vital sign values measured by more than one technique (e.g., invasive/non-invasive blood pressures) were grouped together and names were standardized (e.g., "heart rate" and "pulse" were both standardized to "heart rate"). Numeric results containing text (e.g., a comment or result interpretation) or invalid characters (e.g., "<", ">") were extracted and then any remaining non-numeric values were removed. Results of '0' were also removed as these typically indicate a bad or invalid value in the EHR system. Finally, duplicate results were removed.

4.1.1.3 Feature Generation

Features were defined separately for non-temporal and temporal data. Non-temporal data included age, sex, race, length of stay, mechanical ventilation information, recorded locations, and

diagnoses. Mechanical ventilation information was used to define a Boolean feature indicating presence or absence of a recorded ventilator event. Recorded locations were used to identify the pediatric ICU admitting unit (defined as the unit location immediately prior to the first recorded visit to the pediatric ICU). From the diagnoses, three features were extracted: 1) flag indicating presence of a cancer diagnosis (based on a pre-defined set of ICD-9 and ICD-10 codes); 2) flag indicating presence of cardiopulmonary resuscitation (CPR) (based on a pre-defined set of ICD-9 and ICD-10 codes for cardiac arrest); and 3) admitting diagnosis ICD-10 code category (e.g., for ICD-10 code C40.10, "malignant neoplasm of short bones of upper limb", code category would be C40, "malignant neoplasms of bone and articular cartilage of limbs"). For the admitting diagnosis ICD-10 code category, pediatric ICU admitting unit, and race features, categories with <30 observations were mapped to an "Other" category to ensure each possible category would be represented in the training dataset. Temporal data in the dataset included physical assessment measurements and laboratory test results collected at irregular time intervals. A fixed set of features was defined to summarize the time-series information. Pupil reaction was the only categorical measurement, and was summarized using five features: 1) first value, 2) most recent value, 3) second most recent value, 4) count of results where one pupil was non-reactive, and 5) count of results where both pupils were nonreactive. Each non-categorical temporal measurement was summarized using 17 features comprised of five point estimates (first, minimum, maximum, second most recent, and most recent values) and three trends (difference, percent change, and slope) between the most recent value and all other point estimates (12 features total). The final feature set included 422 features and is described in Table 4.

Missing values were present within the feature set as not all encounters had measurements required to compute each feature. For categorical data, missing values were retained by simply

adding a "missing" category. For numerical data, the data were first discretized using the minimum description length criterion discretization method,[101] which accounts for class information (e.g., in-hospital mortality status) when defining discretization bins. Missing values were then retained by including a "missing" category along with the discretized bins. All features were one-hot encoded prior to learning models.

Table 4. Feature names and definitions

Non-temporal features	
Feature Name	*Definition*
Age	Patient age in days
Sex	Patient sex
Race	Patient race
Length of stay	Elapsed time between arrival date and time of prediction
Pediatric ICU admitting unit	Unit location immediately prior to first recorded visit to pediatric ICU
Admitting diagnosis category	ICD-10 category of admitting diagnosis code
CPR flag	Presence/absence of pre-defined cardiac arrest diagnosis code
Cancer flag	Presence/absence of pre-defined cancer diagnosis code
Mechanical ventilation flag	Presence/absence of recorded ventilator event
Temporal features	
Feature Name	*Definition*
First value	Result with earliest timestamp in defined time-window (missing if <3 results)
Second most recent value	Result with second most recent timestamp in defined time-window (missing if <2 results)
Most recent value	Result with most recent timestamp in defined time-window
Min value	Minimum result recorded in defined time-window
Max value	Maximum result recorded in defined time-window
Change from previous	Most recent value – second most recent value
Change from min	Most recent value –min value
Change from max	Most recent value – max value
Change from first	Most recent value – first value
% change from previous	(Most recent value – second most recent value)/(second most recent value)*100
% change from min	(Most recent value – min value)/(min value)*100
% change from max	(Most recent value – max value)/(max value)*100
% change from first	(Most recent value – first value)/(first value)*100
Rate of change from previous	Slope between most recent value and second most recent value
Rate of change from min	Slope between most recent value and min value
Rate of change from max	Slope between most recent value and max value
Rate of change from first	Slope between most recent value and first value
# results w/ both pupils nonreactive	Count of pupil reaction results where both pupils were nonreactive
# results w/ one pupil nonreactive	Count of pupil reaction results where one pupil was nonreactive

4.1.1.4 Model Learning and Evaluation

To learn and evaluate models, the dataset was split into a training dataset (encounters from 2015) and a test dataset (encounters from 2016). This split was used to simulate model performance when deployed into practice, where the model would be trained on prior years of data and be used to make predictions on future years of data that might include substantial differences from data of prior years. The training dataset was used to perform feature selection techniques and train models, while the test dataset was used to evaluate the models. Two popular strategies for feature selection were examined: 1) correlation-based feature subset (CFS) selection[102], which aims to find a set of features that have high-correlation with in-hospital mortality but low inter-correlation with each other—that is, a set of non-redundant, highly informative features—and 2) information gain (IG) filter with a threshold of 0, which results in selecting features that contain at least some predictive information for in-hospital mortality. CFS feature selection was carried out using the WEKA (Waikato Environment for Knowledge Acquisition) version 3.9.3 implementation[103,104] via the Python package python-weka-wrapper3 version 0.1.7.[105] IG feature selection was carried out using the Python package scikit-learn version 0.20.2.[106]

Several different models were trained, including a Logistic Regression model, which is the standard model utilized in the clinical domain, as well as three frequently utilized ML models—Random Forest, Naïve Bayes, and SVM. Brief overviews of these algorithms can be found in Meyfroidt et al.[107] All models were learned using algorithm implementations provided in the Python package scikit-learn version 0.20.2.[106] Default algorithm settings were adopted for all algorithms, with the exception of the Random Forest model, which was learned using 100 trees instead of the default of 10 trees to improve the performance of the classifier.

Predictive performance of the models was evaluated using the test dataset. Model discrimination was assessed by calculating the area under the receiver operating characteristic curve (AUROC) and 95% confidence intervals (CIs). Due to the large class imbalance in the dataset (only 2% of encounters were death cases), predictive performance was also assessed by calculating the area under the precision-recall curve (AUPRC). The AUPRC is an informative predictive measure that complements the AUROC for imbalanced datasets, i.e., datasets where the outcome of interest occurs rarely.[108] All analyses were performed using R version 3.5.0.[109] AUROCs and 95% CIs were calculated using the pROC package 1.15.3[110] and AUPRCs were calculated using the PRROC package version 1.3.1.[111]

4.1.2 Results

The final dataset included 4,910 encounters (93 in-hospital deaths; 4,817 controls; 1.9% in-hospital mortality rate). The training and test datasets comprised 2,480 (42 in-hospital deaths; 2,438 controls; 1.7% in-hospital mortality rate) and 2,430 encounters (51 in-hospital deaths; 2,379 controls; 2.0% in-hospital mortality rate), respectively. A total of eight models were learned, comprising each combination of feature selection technique and model type (Table 5). Model performance measured by AUROC and AUPRC was comparable for all models, with the exception of the Naïve Bayes model using the IG feature selection approach, which had a very low AUPRC. The Random Forest model using the IG feature selection approach was the highest performing model when examining both AUROC and AUPRC, and thus was selected as the model for which explanations would be designed.

Table 5. Model descriptions and performances

Feature Selection (# features)	Model	AUROC [95% CI]	AUPRC
IG (146)	Logistic regression	0.92 [0.86-0.97]	0.77
	Naïve Bayes	0.92 [0.87-0.96]	0.19
	Random Forest	**0.94 [0.90-0.99]**	**0.78**
	SVM	0.93 [0.87-0.98]	0.78
CFS (8)	Logistic regression	0.94 [0.89-0.98]	0.76
	Naïve Bayes	0.94 [0.90-0.97]	0.74
	Random Forest	0.93 [0.88-0.98]	0.75
	SVM	0.94 [0.89-0.98]	0.73

4.2 Defining Context of Use and Identifying Explanation Design Requirements

In this section, I applied the proposed framework to define an initial context of use and identify promising explanation design requirements for the pediatric ICU in-hospital mortality risk model. All insights are derived from my prior experiences in developing predictive models as well as from an informal review of the literature on interpretable ML, social science work on human explanation and medical decision-making, HCI, information visualization, CDSS (specifically barriers, facilitators, and provider perceptions), and predictive models evaluated by providers or implemented in practice.

I focused specifically on using the predictive model as a tool to support clinical decision-making in the critical care setting by serving as a proxy measure for deteriorating clinical acuity. The end goal of such a system would be to impact critical care provider decision-making and improve clinical outcomes; however, when designing explanations, the more immediate goal would be to promote system adoption. It has been shown that adoption of predictive models can be influenced by provider perceptions of the model utility, credibility, and usability.[112] Thus, it is useful to consider how the framework might inform explanation designs that positively influence these perceptions of the system. For the purpose of this discussion, I defined utility as the perceived

benefit or usefulness of the model (i.e., whether providers can extract meaningful or actionable information from the model), credibility as the "believability" or "persuasiveness" of the model (i.e., whether the model predictions and reasoning processes seem unbiased and aligned with domain knowledge), and "usability" as the feasibility of using the model as part of clinical practice (i.e., ease with which the model can be understood and integrated with existing workflows).

In sections 4.2.1 and 4.2.2, I discuss the target questions in the framework, building upon answers to prior questions when appropriate. For each target question, I address the current understanding of the question based on the available literature and highlight gaps in knowledge that need to be addressed in studies with the target users. Figure 9 provides a summary of the insights for each target question and serves as a guide for the discussions sections 4.2.1 and 4.2.2. In section 4.2.3 I provide a brief commentary on how the answers to these target questions can inform explanation designs that positively influence provider perceptions of the utility, credibility, and usability of the pediatric ICU in-hospital mortality risk model.

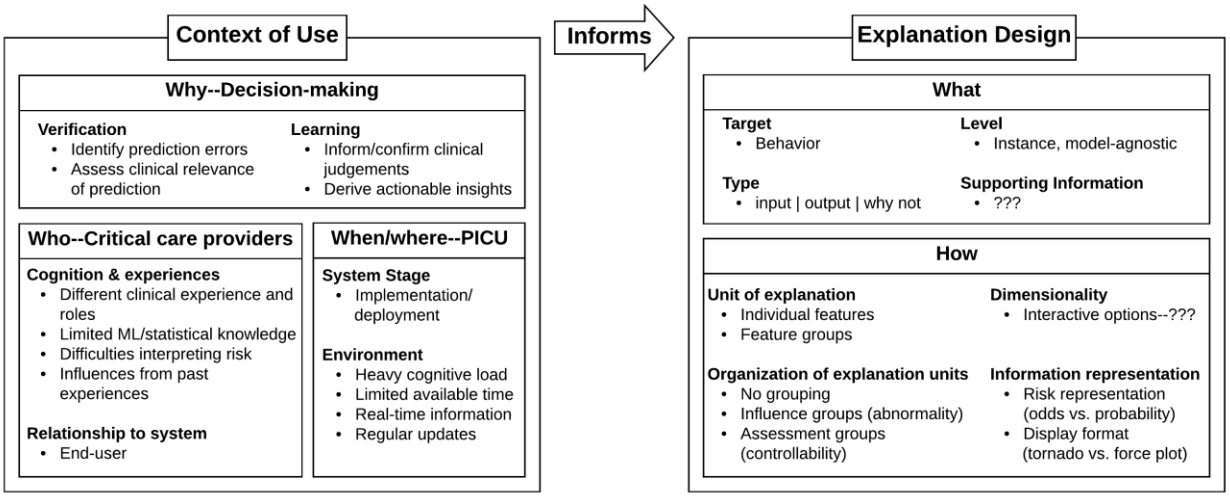


Figure 9. Summary of an initial context of use and a possible space of explanation designs for the pediatric ICU in-hospital mortality risk model. All insights were derived from the literature and prior experiences in developing predictive models for healthcare.

4.2.1 Context of Use

In this section, I define an initial context of use for explanations for the pediatric ICU in-hospital mortality risk model by summarizing the current understanding of *who* might need an explanation, *when* and *where* they might require that explanation, and *why* they want the explanation. I recognize that prior experiences and the literature will not provide a complete picture of the context of use, and thus highlight gaps in knowledge that need to be addressed in approaches involving the identified target users.

Who

Current understanding: The target users for the model are critical care providers. Any member of a critical care team (e.g., nurses, residents, fellows, attending physicians, etc.) would be interested in deteriorating clinical acuity of a patient, and thus might find the predictive model of use. For decision-making in the clinical setting, the relationship of all providers to the model at the time of explanation would be that of an end-user of the system (as opposed to a designer if the scenario of interest was in soliciting expert feedback for model improvement). In terms of user cognition, critical care providers will fall under Ribera and Lapedriza's "domain expert" category[42] and they will typically lack the knowledge to understand and critically evaluate ML models for use in practice.[1,5] Moreover, there is evidence in the literature that providers have difficulties in interpreting risk and probability-based estimates,[12,96,112,113] which suggests that providers might also struggle to understand and evaluate prediction models that employ traditional statistical approaches (e.g., logistic regression models).

In addition to user cognition, past experiences of a user could influence their explanation needs and design requirements. For example, negative experiences with past predictive models or

health information technology may lead users to require that an explanation include specific information or be designed in a specific manner to prevent recurrence of past experiences. Examples of negative experiences with health information technology and predictive models in the literature include: 1) inappropriate or disruptive alerts,[93,97,112,114,115] 2) high effort to use the system,[35,37,45,112] 3) information that lacks clinical utility (e.g., incorrect, irrelevant, not actionable),[36,45,46,112] and 4) lack of control.[36,112]

Gaps in knowledge: The level of statistical and ML knowledge of critical care providers is unknown, although the literature suggests that most providers will have a limited understanding of the topics. It has been shown that critical care providers employ different information seeking strategies based on their clinical training and role in the patient care process,[116] which suggests that users with different clinical positions and knowledge may require different explanations for the same predictive model. These possible differences require further exploration in reference to the target users. Although the literature highlights some possible negative experiences that critical care providers may have previously had with health information technology and predictive models, the influence of past experiences will vary by user and setting and again requires further exploration regarding target users.

When and where

Current understanding: For use in clinical decision-making where critical care providers are end-users of the predictive model, explanations will be provided at the deployment/implementation stage of the system. For the predictive model, this would constitute use of the explanation in the pediatric ICU, which is a complex, dynamic environment where information is abundant and decisions are time-sensitive. To be useful in clinical decision-making

in this environment, a predictive model must be able to keep up with the influx of data to provide real-time predictions (and explanations)[3,12,34] and be regularly updated to reflect changes in patient populations and care processes.[89,117] Additionally, evidence suggests that a successful tool should support existing clinical workflows,[35,37,45,112] such as respecting a provider's available time and current cognitive load. More specifically, the tool (and by extension its explanations) should avoid contributing to information overload[118,119] or requiring large time investments to use (e.g., manual data entry).[35,45,96,115]

Gaps in knowledge: As workflow fit is an important factor of successful adoption, further exploration of the pediatric ICU workflow and environment is required.

Why

Current understanding: To use the predictive model in decision-making, a healthcare provider must be able to integrate the model information with their knowledge, experience, and missing contextual information and then translate the information into a meaningful decision.[3,45,93,97] This process usually occurs at the patient-level (i.e. for an individual prediction), and involves the closely related goals of verification and learning. In verification, providers assess a prediction to determine if it is clinically relevant (i.e., aligns with domain knowledge) before using it to inform clinical decisions. Verification is especially important in the context of ML, as models that perform well on average are often based on statistical associations and imperfect data,[3,12] may be missing important information about a patient (e.g. contextual information),[45,113] and may therefore have significant individual level errors.[1,120] Providers must be able to understand model limitations and identify errors to determine whether a risk prediction applies to a specific patient and defend any decisions based on the prediction.[1,3,6,12,97,121] When learning from

a predictive model system, providers extract knowledge from the system to assist in decision-making, which may include informing or confirming clinical judgments[95] and deriving actionable insights (e.g., identifying potentially modifiable risk factors).[7,96,112] Verification and learning often occur simultaneously, such as when providers investigate discrepancies in the match between their knowledge and the model's knowledge (i.e., a prediction seems too high or low). They may start by looking for a source of model error (verification) but end up discovering a risk factor that was overlooked (learning).

Gaps in knowledge: Verification of and learning from individual predictions are assumed to be the main explanation goals for critical care providers using the model in decision-making, but it is worth verifying this assumption with the target users. Additionally, discussions with target users can help elucidate the information they require to verify model predictions as well as determining what information they would be interested in learning from the model.

4.2.2 Explanation Design Requirements

In this section, I utilize the defined context of use and insights from prior experiences and the literature to suggest promising design requirements for *what* information the explanation needs to contain and *how* to provide that information to target users. I recognize that this approach will not clearly identify all explanation design requirements, and thus highlight gaps in knowledge that need to be addressed in approaches involving target users.

What

Possible explanation design requirements: The target user goals are verification and learning, which the literature suggests can be supported by showing the influence of risk factors

on a specific prediction, as this facilitates comparison with clinical knowledge and can assist in deriving actionable insights.[45,93,96,97,121] This finding suggests that target users are likely seeking contrastive explanations ("why not" type explanations) for specific predictions, or explanations that demonstrate which inputs are pushing the prediction toward one outcome over another. As the target users are expected to have limited ML knowledge, explanations targeted at explaining model behavior, or relationships between inputs and outputs, are likely appropriate. Moreover, as it is expected that the target users will use individual predictions to assist in decision-making, it also seems appropriate to provide explanations at the instance-level (i.e. patient-level explanations). Prior work has supported the use of instance-level explanations for predictive models in healthcare as they provide insights on individual patients and thus support precision medicine initiatives.[10,16]

The aforementioned design requirements can be met by existing post-hoc explanation approaches that provide instance-level explanations based on feature influence values. Utilizing a model-agnostic approach to explanation would provide further benefit, as the environment of use requires that the explanations place limited burdens on cognitive load and processing time and that the predictive model be continually updated over time. A model-agnostic explanation approach would allow the explanation design to be tailored to reduce cognitive load and processing time without having any impact on the underlying predictive model or its accuracy. Moreover, a model-agnostic explanation approach allows the predictive model to adapt over time with minimal changes to an explanation design familiar to providers.

Gaps in knowledge: Although the context of use and literature insights support the use of model-agnostic, instance-level, explanation approaches based on feature influence methods, the utility of these explanations has not been verified in studies with healthcare providers. It is also unclear from the literature what supporting information critical care providers might need to

understand these explanations. Inquires with target users are required to validate the appropriateness of these explanations and understand what supporting information facilitates target user understanding of the explanation. Moreover, it is unclear from the literature whether "why not" type explanations will be sufficient for the target users; other types of explanations may be required (e.g., "what if" explanations that allow providers to simulate how a change in a feature affects a patient's prediction).

How

Identified design requirements: As model-agnostic, instance-level, explanation approaches based on feature influence methods were identified as a promising explanation approach, design options discussed in this section relate to the presentation of these types of explanations. As the target users are expected to have limited understanding of ML and will be using explanations in a cognitively demanding and time-constrained environment, explanation content should be presented in a manner that facilitates information processing with minimal demands on cognition and time. When considering the unit of explanation, utilizing larger cognitive chunks can reduce the cognitive load and processing time required. For instance-level explanations of feature influence, a larger cognitive chunk could be obtained by grouping features by laboratory test or vital sign instead of showing the individual features derived for each test or vital sign. There is some evidence supporting the use of feature groupings and high-level feature abstractions for non AI/ML experts.[63]

Organizing explanation units into meaningful groups can also potentially reduce the cognitive load and processing time. For explanations based on feature influence, the standard organization of explanation units would simply be a list of the features in decreasing magnitude of

influence on the prediction (i.e. a ranked list). However, there is some evidence that users prefer meaningful groupings of explanation units over ranked lists.[122] Social science literature indicates that humans prefer explanations that include causes that are abnormal or controllable (i.e., modifiable),[27] suggesting that grouping explanation units by these factors might prove more informative. For instance-level explanations of feature influence, grouping by abnormality could be achieved by grouping features by whether they increase or decrease the predicted risk. Grouping by controllability or modifiability could be simulated by grouping features by whether they are static (i.e., cannot be changed through intervention such as age) or dynamic (i.e., could be changed through intervention such as a laboratory test result).

Reducing the dimensionality of an explanation can also lead to decreased demands on processing time and cognitive load, but must be balanced with a user's ability to understand a prediction. Dimensionality refers to the level of detail of the explanation, which can be reduced through information removal (e.g., reducing explanation size) or aggregation (e.g., reducing explanation granularity). There is some evidence in the literature that the desired dimensionality of an explanation will vary by individual and prediction, i.e., some users prefer more detailed explanations and users often want more detailed explanations for high risk predictions,[68,122] which suggests that controlling dimensionality via interactive options may be beneficial. This aligns with concepts from social science literature that explanation should occur as part of a conversation, where users may ask for additional information or explanations after receiving the initial explanation.[27,56] For the model with instance-level explanations of feature influence, interactive options for controlling dimensionality could include control over the granularity of the units of explanation (e.g., whether to view individual features or feature groups), control over how units of

explanation are organized or grouped (e.g., by increasing/decreasing risk), and/or control over the size of the explanation (number of units of explanation shown).

Finally, the vocabulary, data structures, and visualizations used to express information can impact how effectively critical care providers can process an explanation. The vocabulary of the explanation should include standard clinical terms familiar to the critical care providers and should represent risk in a format with which critical care providers are comfortable. The literature suggests that visual or graphical representations of risk information can facilitate healthcare provider comprehension of risk, but this has not been validated in user studies.[96] For instance-level explanations of feature influence, it should be clear to critical care providers how each feature contributes to the predicted risk. For feature influence explanations, feature contributions to the predicted risk have been previously represented in terms of odds or probability (e.g., a feature increases risk 2-fold or by 10%) and visualized using tornado plots and custom visualizations called force plots (see Figure 12 in section 4.3 for an example).[75,123]

Gaps in knowledge: Overall, the context of use and literature provide some possible design options that might prove beneficial when presenting instance-level explanations of feature influence for predictions from the model. However, discussions and feedback from the target users are required to further understand how critical care provider cognitive load and processing time might be affected by 1) what units of explanation are used (e.g., feature groupings or individual features), 2) how the units of explanation are organized (e.g., no grouping, grouping by increasing/decreasing risk, grouping by dynamic/static), 3) what interactive options are provided for controlling explanation dimensionality, and 4) what vocabulary, data structures, and visualizations are used to present the predicted risk, the explanation, and any supporting information.

4.2.3 Potential impact on perceptions

As discussed in section 4.2.1, explanations for the pediatric ICU in-hospital mortality risk model would likely be used by critical care providers who: 1) have limited knowledge of statistical and ML concepts (*who*), 2) work in a cognitively demanding and time-constrained environment (*when/where*), and 3) would be seeking explanations to assist them in verifying predictions from the model and learning information that can assist in decision-making (*why*). Therefore, as discussed in section 4.2.2, the explanation design should contain (*what*): 1) content that supports a provider's current information goals (e.g., verification and learning), which would likely increase the perceived utility and credibility of the model; and 2) appropriate supporting information to help a provider interpret the explanation, which would likely increase the perceived usability of the predictive model. Moreover, the explanation design should present information (*how*) in a manner that reduces the cognitive load and processing time required by a provider, which would likely increase the perceived usability of the model.

4.3 Preliminary Explanation Designs

As noted in section 4.2.2, there are several aspects of the explanation design that require further investigation in discussions with target users. To facilitate these discussions, I proposed preliminary explanation designs for the pediatric ICU in-hospital mortality risk model. As per the insights from section 4.2, I focused on suggesting explanation designs for model-agnostic, instance-level explanations based on feature influence methods to better understand the potential utility of these types of explanations within the healthcare domain.

Two popular and publicly-available model-agnostic, instance-level explanation algorithms have been previously applied to predictive modeling problems in the healthcare domain—the local interpretable model-agnostic explanations (LIME) algorithm[64,123] and the Shapley additive explanations (SHAP) algorithm.[124,125] The LIME algorithm generates an explanation for a prediction by learning an interpretable model (e.g., sparse linear regression) that fits the local decision boundary near the instance of interest. The SHAP algorithm is based on concepts from game theory and is theoretically guaranteed to be faithful to the underlying predictive model. It unifies several alternative instance-level explanation algorithms into a single approach, including the LIME algorithm. A detailed description of both algorithms is available in Appendix A. To generate model-agnostic, instance-level explanations of feature influence for the pediatric ICU in-hospital mortality risk model, the SHAP algorithm was used after a series of experiments comparing the two algorithms revealed that the LIME algorithm did not guarantee local fidelity and required more computation time. The experiments are described in Appendix A.

Based on insights from section 4.2, I mocked-up five explanation designs for the SHAP explanations to solicit critical care provider feedback on the following explanation design options:

1.) Unit of explanation—individual features (low granularity) vs. feature groupings by lab test/vital sign (high granularity)
2.) Organization of explanation units—no groupings, grouping by influence on risk (i.e. whether the unit increases/decreases risk), grouping by assessment (e.g., laboratory test features, physical assessment test features, demographic/healthcare utilization features) which was used as an approximation of the controllability of features
3.) Dimensionality—static vs. modifiable explanation (i.e., interactive options to control explanation size and granularity of explanation unit)

4.) Risk representation—probability vs. odds

5.) Explanation display format—tornado plot vs. force plot

Each mock-up included the predicted risk of mortality from the pediatric ICU in-hospital mortality risk model, an explanation for the predicted risk from the SHAP algorithm, and supporting information to assist in interpreting the risk and explanation. Mock-ups varied in explanation design options and were organized into two sets based on the different design options. The mock-up sets are summarized in Table 6 and the explanations for each mock-up are shown in Figures 10-14. By default, mock-ups with feature groups for the unit of explanation included groupings by influence within the explanation plot (i.e., each feature group had factors that increased or decreased the risk). Mock-ups with feature groups and tornado plots also included an interactive hover-box option to view the individual level features within each group (i.e., modifiable granularity of explanation unit). For mock-ups with modifiable explanation size, an interactive option to scroll down the explanation plot to view additional features was included.

Table 6. Explanation design options used for each mock-up

			Set 1			Set 2	
			1-1	**1-2**	**1-3**	**2-1**	**2-2**
Unit of explanation	Individual features		X			X	
	Feature groups			X	X		X
Organization of explanation units	None		X				
	Influence groups			X	X	X	X
	Assessment groups						X
Dimensionality	Size	Static			X		
		Modifiable	X	X		X	X
	Granularity of explanation unit	Static	X		X	X	
		Modifiable		X			X
Risk representation	Probability			X	X	X	X
	Odds		X				
Explanation display format	Force plot				X		
	Tornado plot		X	X		X	X

Supporting information for each mock-up included demographic information (e.g., age, length of stay), a list of current diagnoses, a table of the raw values of the features used in the model (i.e., undiscretized feature values), and an interactive plot where the raw values of time series data from laboratory tests and vital signs could be viewed. An example of the supporting information included in each mock-up is shown in Figure 15. SHAP explanations were generated using the Python package shap version 0.27.0[126] and mock-ups were generated as interactive HTML pages using the Python package bokeh version 1.0.4.[127]

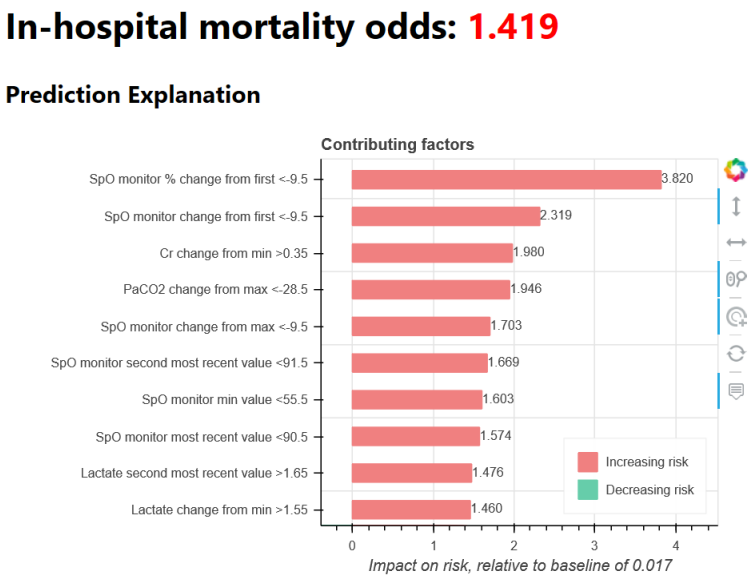


Figure 10. Mock-up 1-1 prediction and explanation. This mock-up depicts the following design options:1) unit of explanation—indiviudal features, 2) organization of explanation units—no grouping, 3) dimensionality—modifiable explanation size and static granularity of explanation unit, 4) risk representation—odds, and 5) explanation display format—tornado plot.

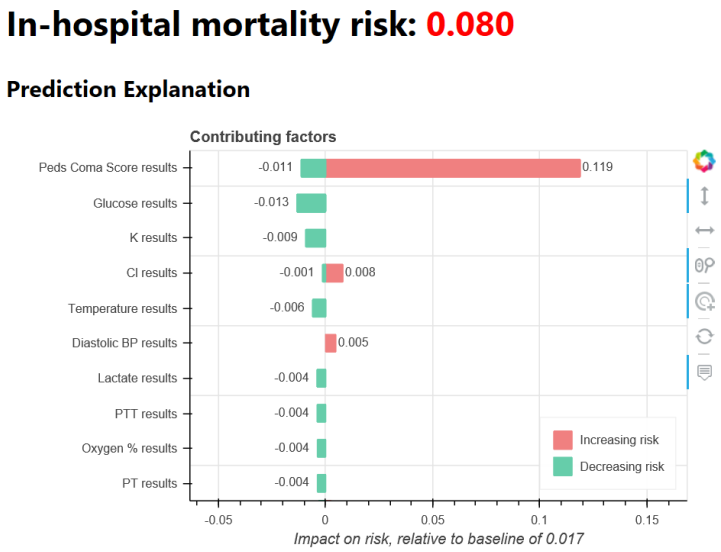


Figure 11. Mock-up 1-2 prediction and explanation. This mock-up depicts the following design options:1) unit of explanation—feature groups, 2) organization of explanation units—influence groups, 3) dimensionality—modifiable explanation size and modifiable granularity of explanation unit, 4) risk representation—probability, and 5) explanation display format—tornado plot.

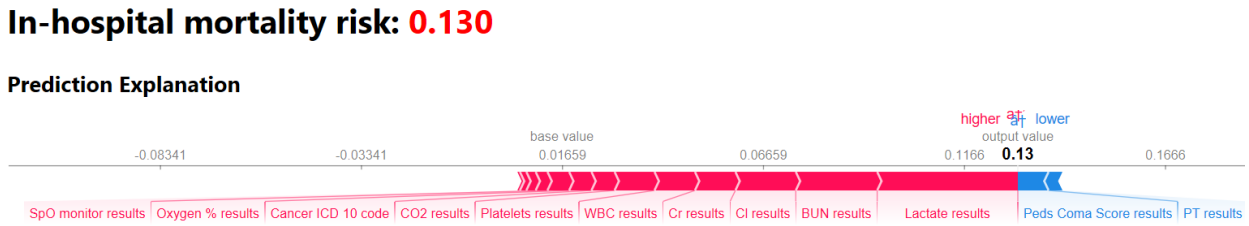


Figure 12. Mock-up 1-3 prediction and explanation. This mock-up depicts the following design options:1) unit of explanation—feature groups, 2) organization of explanation units—influence groups, 3) dimensionality—static explanation size and static granularity of explanation unit, 4) risk representation—probability, and 5) explanation display format—force plot.

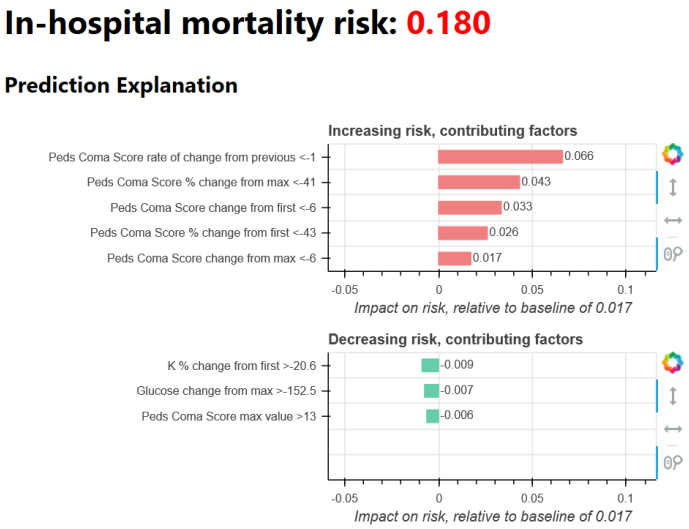


Figure 13. Mock-up 2-1 prediction and explanation. This mock-up depicts the following design options:1) unit of explanation—individual features, 2) organization of explanation units—influence groups, 3) dimensionality—modifiable explanation size and static granularity of explanation unit, 4) risk repsentation—probability, and 5) explanation display format—tornado plot.

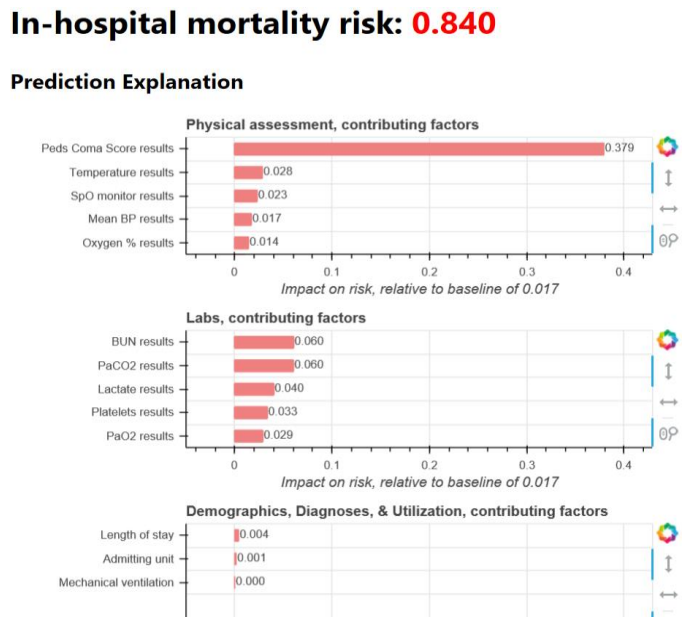


Figure 14. Mock-up 2-2 prediction and explanation. This mock-up depicts the following design options:1) unit of explanation—feature groups, 2) organization of explanation units—influence groups and assessment groups, 3) dimensionality—modifiable explanation size and modifiable granularity of explanation unit, 4) risk repsentation—probability, and 5) explanation display format—tornado plot.

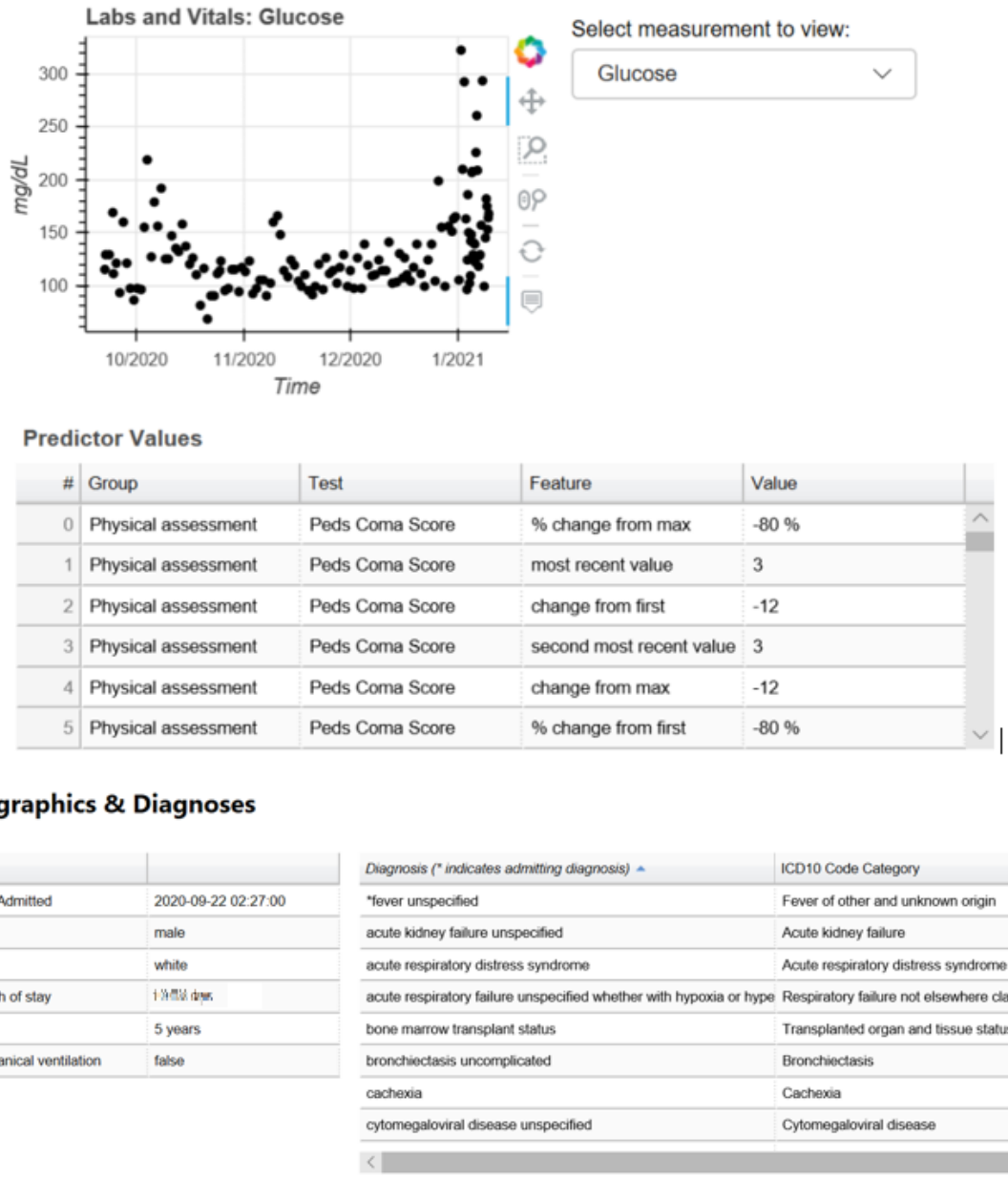


Figure 15. Supporting information provided in each mock-up. Each mock-up included included demographic information (bottom left), a list of current diagnoses (bottom right), a table of the raw values of the features used in the model (middle) and an interactive plot where the raw values of time series data from laboratory tests and vital signs could be viewed (top).

5.0 User Studies to Refine Explanation Design

I conducted focus groups with critical care providers to refine the defined context of use and solicit feedback on the mock-ups of the explanation designs for the pediatric ICU in-hospital mortality risk model proposed in section 4.3. More specifically, I aimed to:

1.) Assess critical care provider attitudes about using the predictive model in practice
2.) Assess critical care provider perceptions of the model-agnostic, instance-level approach for explaining predictions from the model
3.) Explore critical care provider preferences on the design options proposed in the mock-ups to identify those that facilitate understanding of and positively influence perceptions of the model

Insights from the focus group were used to inform a final user-centered explanation design to be used in a laboratory study to evaluate the impact of a user-centered explanation design on critical care provider decision-making and perceptions of the pediatric ICU in-hospital mortality risk model. Section 5.1 describes the materials and methods of the study, section 5.2 summarizes the main results, and section 5.3 presents the final user-centered explanation design.

5.1 Materials and Methods

5.1.1 Setting and Participants

All focus groups were conducted at CHP during March 2019-June 2019. A convenience sample of pediatric critical care providers of differing clinical expertise (e.g., nurses, residents, fellows, attending physicians) was recruited through professional connections of one of the dissertation committee members to participate in focus group sessions. Participants were assigned to sessions based on availability. The study was approved by the University of Pittsburgh IRB.

5.1.2 Procedures and Data Collection

I conducted a total of three focus group sessions, each ~1.5 hr in length and comprising 5-8 participants. Each participant attended only a single focus group session, during which they were asked to participate in four activities:

1) *Background questionnaire (~5 mins):* Participants were asked to complete a background questionnaire assessing their clinical experience, familiarity with predictive modeling, and perceptions of predictive analytics.
2) *Model discussion (~30 mins):* Participants listened to a presentation on the development of the pediatric ICU in-hospital mortality risk model and then participated in a guided group discussion about their initial perceptions of the model.
3) *Mock-up review discussion (~50 mins):* Participants received brief training on interpreting explanation information and then participated in a guided group review and critique of the five mock-ups of explanation designs for predictions from the

pediatric ICU in-hospital mortality risk model. Mock-ups were reviewed by set. For each set, mock-ups were discussed individually and then presented side-by-side to facilitate discussions of preferences for the different design options. Participants were provided with print-outs of each mock-up and encouraged to write comments and design suggestions on the sheets.

4) *Ranking questionnaire (~5 min):* Participants were asked to complete a questionnaire to indicate their preferred variation of each design option presented in the mock-ups and to rank the options in order of perceived importance in understanding the model prediction.

A focus group script and question guides were developed and followed for each session. Copies of the background questionnaire, question guides for the discussion activities, and the ranking questionnaire can be found in Appendix B. Focus group sessions were moderated by 1-2 researchers and a separate researcher took notes during each session. All sessions were audio-recorded and all materials (questionnaires, print-outs of mock-ups) were collected from participants at the end of each session.

5.1.3 Data Analysis

Background questionnaire and ranking questionnaire responses were summarized using descriptive statistics and visualizations. Audio recordings of the sessions were transcribed verbatim and written participant comments on mock-up print-outs were compiled by session. Transcripts and written participant comments were analyzed using descriptive coding.[128] One analyst developed an initial codebook with the concepts and definitions from the proposed framework along with codes to capture participant perceptions of the utility, credibility, and

usability of the system. The analyst then applied the codes to the transcripts and written participant comments, refining definitions and adding codes to more finely represent the participants' responses. A second analyst used the codebook to independently code one session transcript. The two analysts discussed coding differences to resolve disagreements and achieve consensus on a final codebook (Appendix C). The first analyst then recoded all transcripts and written participant comments. QSR International's NVivo 12 software[129] was used to assign and organize codes. Session notes recorded by the researchers were not coded, but were used to assist in coding and interpretation. This analysis was intended to identify insights related to each of the target questions in the proposed framework to address the gaps in knowledge identified in section 4.2. Insights from the coding process were analyzed in conjunction with questionnaire responses to summarize findings about the context of use and explanation design and identify elements that influence critical care provider perceptions of the pediatric ICU in-hospital mortality risk model.

5.2 Results

A total of 21 critical care providers participated in the three focus group sessions. Table 7 summarizes the clinical experience of the participants in each session. The following sections summarize insights on the context of use and explanation design for the pediatric ICU in-hospital mortality risk model, specifically highlighting factors that influenced perceptions of the model.

Table 7. Summary of participants in each focus group

Session	# of Participants	Clinical Experience		
		Attending	Fellow/resident	Nurse
1	5	3	2	0
2	8	6	2	0
3	8	0	0	8

5.2.1 Insights on Context of Use

Table 8 provides a high-level summary of the ideas discussed in this section, specifically major insights related to the context of use and elements that would influence perceptions of the model credibility, utility, and usability. Table 9 and Figure 16 summarize the participants' background knowledge and attitudes towards predictive modeling, respectively. Insights identified for each of the target questions related to context of use (*who*, *when/where*, *why*) are summarized with supporting quotes in Table 10. Although the insights are separated out by target question in Table 10, I summarize the findings about the context of use as a whole.

Table 8. High-level summary of insights on context of use and influences on perceptions of the model

<table>
<tr><th colspan="2"></th><th colspan="2">User characteristic (who)</th><th>Desired information</th><th>Factors that would positively (+) or negatively (-) influence perceptions</th></tr>
<tr><td rowspan="3">Explanation goal (why)</td><td rowspan="2">Verification</td><td rowspan="2">Predictive modeling knowledge</td><td>Detailed</td><td>• Predictive performance
• Alignment with domain knowledge
• Comparison with existing models
• Modeling processes</td><td rowspan="2">Credibility
+ high predictive performance
+ predictions that aligned with clinical knowledge
- influential outliers or data errors
- counterintuitive risk factors
- model limitations</td></tr>
<tr><td>Basic</td><td>• Predictive performance
• Alignment with domain knowledge</td></tr>
<tr><td>Learning</td><td>Clinical role</td><td>Physician</td><td>Obtain insights about patients
• Prioritization
• Assessment of status
• Highlight patients/info of concern</td><td>Utility
- insufficient training for users
- clinically irrelevant information

Usability
+ clinically appropriate alerts
- high cognitive effort or attention
- large time investments</td></tr>
<tr><td colspan="2"></td><td colspan="2">Nurse</td><td>Actionable information
• Alerts to important changes
• Information to intervene or justify request for consult</td></tr>
</table>

Participants exhibited wide variation in predictive modeling knowledge (Table 9; Table 10, User Cognition), which affected the types of information providers wanted in order to assess the credibility of the predictive model (Table 10, Verification). All providers compared the model information against domain knowledge; however, providers with more detailed knowledge of predictive modeling also wanted information about the development process (e.g., cohort

definition, data collection and cleaning procedures) and how the model compared to similar, existing models. Perceptions of model credibility were positively influenced by the high predictive performance and model predictions that aligned with domain knowledge (i.e., the participant could clinically rationalize why the model made the prediction). Perceptions of model credibility were negatively influenced by limitations in the modelling process (e.g., not accounting for feature correlations) and model information that did not align with clinical knowledge (e.g., errors in data values, predictions based on outliers or with counterintuitive risk factors).

Table 9. Summary of participant background knowledge of predictive modeling concepts

Familiarity with risk prediction models (1 participant left question blank)	
	# of participants (n=20)
"I know what a risk prediction model is."	17
"I have used a risk prediction model in practice."	8
"I have been involved in the development of a risk prediction model."	6
Familiarity with machine learning (1 participant left question blank)	
	# of participants (n=20)
None—I have never heard of this term before.	5
Basic awareness—I have heard of the term, but don't know much about it.	6
Know a little—I am familiar with the main concepts of machine learning.	4
Know a fair amount—I have a practical understanding of machine learning concepts.	3
Know it well—I have a theoretical understanding of machine learning concepts.	2

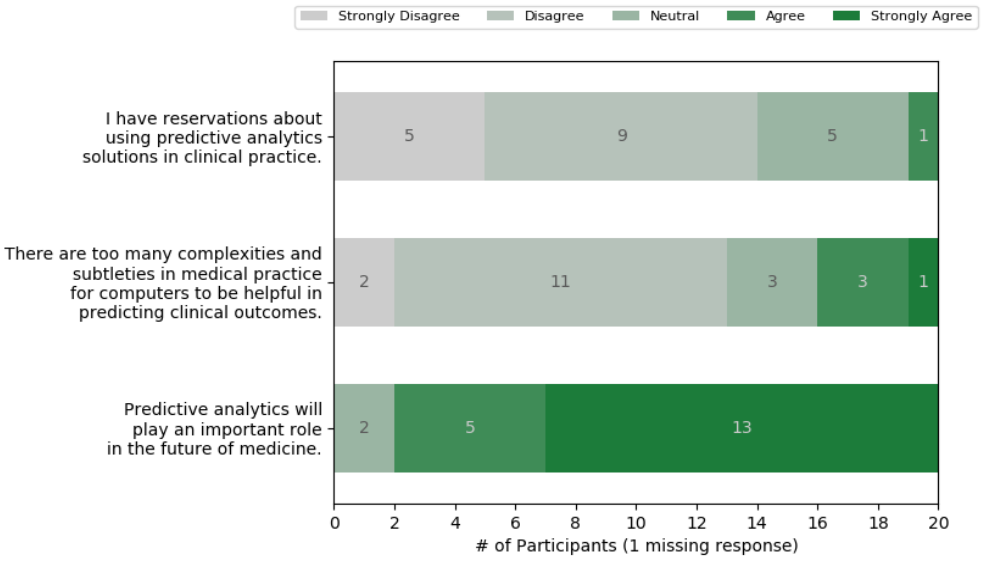


Figure 16. Participant attitudes towards predictive analytics

Participants generally had positive attitudes towards using predictive analytics in clinical practice (Figure 16), and saw several possible applications for information from the pediatric ICU in-hospital mortality risk model. The information participants sought from the model depended on their clinical role (Table 10, User cognition) and factors related to their work environment (Table 10, Social and organizational influences). Physicians anticipated using the model to facilitate patient prioritization during rounds and sought to gain insights about the condition of patients (Table 10, Learning). They viewed the model as a useful tool to help them synthesize patient information and alert them to high-risk patients and/or concerning information. Nurses anticipated using the tool to assist in communicating changes in patient conditions to physicians and generally sought actionable information from the model. They wanted to be alerted to clinically important changes in patient condition and be given information to either intervene or justify a request for a physician consult. However, prior experiences with alerting systems raised concerns from the nurses about appropriate timing and relevance of alerts (Table 10, User cognition). Regardless of clinical role, information seen as clinically irrelevant (e.g., a high risk prediction driven by a low Coma score for a sedated and paralyzed patient) contributed to negative perceptions of model utility. Information that would be difficult to process within the constraints of the ICU environment (Table 10, Cognitive and time resources) also contributed to negative perceptions of usability.

Participants did not generally seek explanations to help them improve the model, but negative perceptions of the model prompted participants to suggest improvements (Table 10, Improvement). These included incorporating additional relevant predictors (e.g., medications), predicting additional outcomes (e.g., morbidity), defining normal ranges for variables (either by age or by setting patient-specific baselines), setting patient-specific alert thresholds, and incorporating domain knowledge into the model.

Table 10. Insights on context of use target questions

	Topic	Insights
Who	User Cognition	*Large variation in knowledge of predictive modeling* "Are we enabling the computer take over?" "Seems like there's a lot of collinearity there. And if those are two of the main contributors, I'm wondering whether it's overestimating" *Negative experiences with prior systems due to insufficient training, irrelevant information, and inappropriate disruptions in workflow* "They trigger a sepsis screen every time I do vitals or every 2 hours and you call the doctors and they have to come down and see them and they're getting really irritated. They don't want to have to do that, they don't have time." "Most people didn't know how to use the Rothman index." *Clinical role affects how providers anticipate using model information* "From the attending perspective, when you walk in in the morning who do you need to see first? Who maybe is higher risk than people are appreciating? Or is changed based on data that's emerged in the last few hours?"
When/where	Cognitive and time resources	*Providers have limited available time to process information* "I don't know if—working on the floor—I get an alert that I have time to go and look through all of this data to try and figure out where the risk is coming from." *Providers have limited available cognitive capacity and attention to process information* "If you load it with a lot of numbers that will probably be not helpful…Because it dilutes your attention." "Trying to think about what that actually means…one, you could be wrong, or if it's 3:00 in the morning…some of the mental gymnastics you'd have to do."
	Social and organizational influences	*Workflow and social factors determine how providers anticipate using model information* "Typically, I round on new patients and then I round on any ECMO patients…and then with some filler between, I break up and it's somewhat random where I start" "I think it might help with the doctors…If there's another number—something saying 'here, look at this, this is actually showing that there's something going on.'….Because I get push back all the time."
Why	Verification	*Providers desired comparisons to existing models and information on model development processes to help validate model* "And is it better than PRISM or is it the same?" "We're assuming that you didn't just select for the sickest patients in this cohort." "Does this weed out their error of charting, things like this?" *Providers validated model information by comparing to domain knowledge* "I mean arterial pressure of 250 seems physiologically impossible." "The leading variables are patient is having respiratory issues and has kidney injury…from a face validity standpoint—yes, that sounds like a patient with a higher risk of dying."
	Improvement	*Providers were interested in improving the performance and utility of the model* "It'd be nice to look at morbidity as well and other things." "One of the critical things that you might consider finding a way to incorporate into the score is medications." "I do think, from a model validity standpoint, changing this to include maybe abnormal blood pressure for age, does add a lot."
	Learning	*Providers wanted to use the model to gain insights about patient conditions* "Does this model offer new information that I didn't already have? Like 'this patient was at high risk for mortality and I didn't otherwise recognize that.'" "Just telling you what you should know, and what you would appreciate if you clicked into the chart and dove into the information, but at least this is synthesizing that for you." *Providers sought actionable information from the model* "Can I do anything to mitigate that risk of mortality based on what I know?" "I don't think there's a lot I can do about most of that stuff…"

5.2.2 Insights on Explanation Design

Table 11 provides a high-level summary of the insights on explanation design discussed in this section, specifically the content to include in the explanation design as well as the preferred design options that would positively influence perceptions of usability. Explanation design was not observed to influence perceptions of credibility or utility. Participant preferences for each design option and the perceived importance rankings of each option are summarized in Figures 17 and 18, respectively. Insights identified for each of the target questions related to explanation design (*what, how*) are summarized with examples quotes in Table 12. Findings about the explanation design are summarized below, referring to insights on the context of use where appropriate.

Table 11. High-level summary of insights on explanation design

Desired content (what)	Benefits	Preferred design options (how)
Explanations: • instance-level, model behavior (SHAP explanations) • global-level, model processes	• Assess model credibility and utility	• Risk expressed as percent probability • Feature groups with details on demand • Interactive options to support different displays/organizations for various users • Familiar icons • Readable from left/right or top/bottom
Table of raw feature values	• Interpret discretized features • Examine trend-based features	• Directionality for trend-based features • Simpler terminology
Time-series data plots	• Investigate suspicious values • Assess trends and baselines	• Multiple plots • Highlight points related to features • Auto-population of data
Contextual information	• Clinically meaningful interpretation	N/A
Risk baselines and trends	• Context for risk prediction	• Prominent display of baseline risk

By providing the plots of the SHAP explanation, the list of predictors that went into the model, and the predicted risk from the model, each mock-up provided "why not"[2], "input", and

[2]SHAP explanations show which inputs are responsible for pushing a prediction toward one outcome over another, which means they are contrastive explanations. This makes them "why not" type explanations.

"output" type explanations, respectively. Participant questions about the model indicated that they were seeking these types of explanations, with some providers also seeking "what if" type explanations, i.e., how the output might change if an input is changed (Table 12, Type, target, and level of explanation). Participants sought all types of explanations when verifying model information (Table 10, Verification), but typically sought "why not" type explanations when seeking information for use in clinical practice (Table 10, Learning). The instance-level explanations of model behavior provided by the SHAP algorithm were generally perceived as helpful in assessing model credibility and utility; however, some providers also requested global-level explanations and explanations of model processes (Table 12, Type, target, and level of explanation). Although useful in assessing credibility and utility, type of explanation was not observed to have a specific influence on participant perceptions of the model.

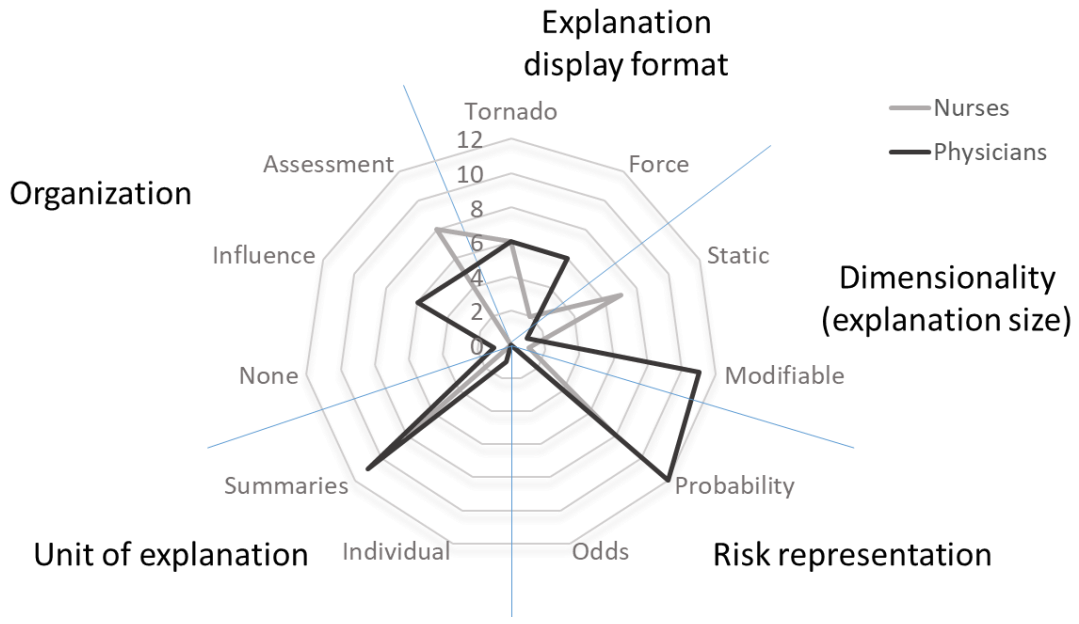


Figure 17. Particpant preferences for design options by clinical role

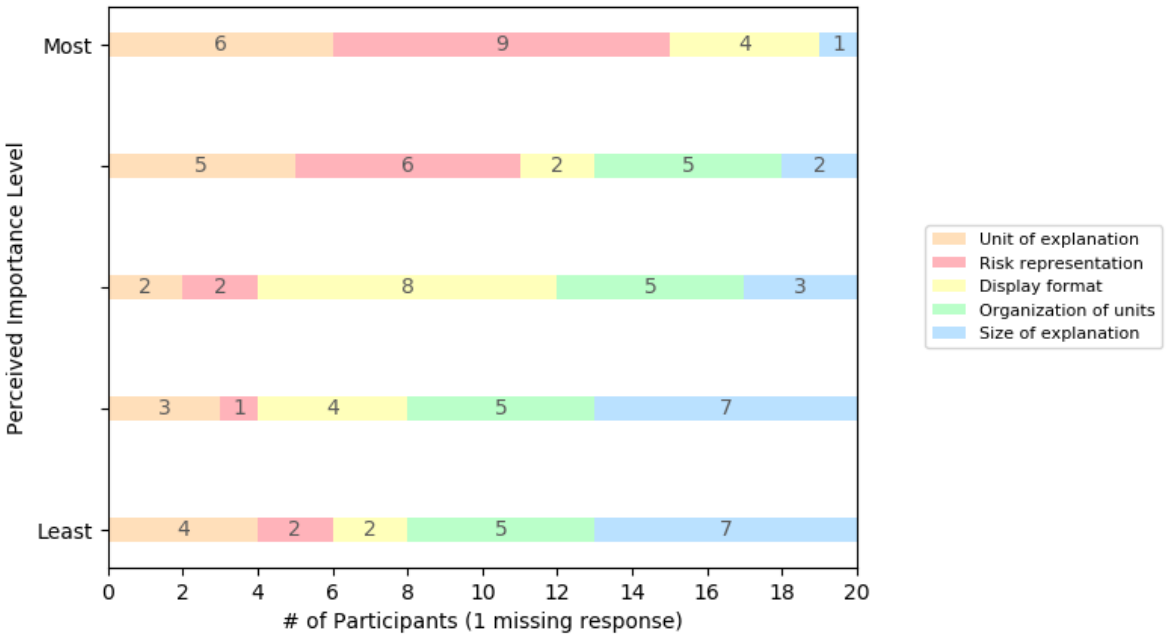


Figure 18. Participant rankings of design options by perceived importance

The supporting information provided in the mock-ups was vital to participant interpretation of the prediction and explanation, and in general helped participants assess the credibility and utility of the model information (Table 10, Verification; Table 10, Learning). Participants frequently utilized the table of the raw values of the features used in the model (i.e., undiscretized feature values), the interactive plot to view the raw values of time series data from laboratory tests and vital signs, and contextual information (e.g., diagnoses) to assist in their interpretation (Table 12, Supporting information). Participants found that the raw values of features used in the model helped them when interpreting the discretized features used in the explanation. For example, if the most predictive feature in the explanation was “Cr change from min >0.35” (i.e. the most recent value of creatinine has increased more than 0.35 since the minimum value), participants found that to assess the clinical importance of the feature it was helpful to also know the exact amount creatinine had increased, as well as what the current and minimum values of creatinine were. Plots

of the raw values of time series data from laboratory tests and vital signs were often utilized to verify model information, specifically to investigate suspicious values (e.g., outliers or errors), assess trend-based features, and determine patient baselines (e.g., if the patient normally has an elevated creatinine level). Contextual information not included in the model (e.g., diagnoses, interventions) was considered essential to assess the clinical validity and relevance of the prediction. For example, when examining a high-risk prediction that included a low coma score as a top predictor, participants noted that they would give less weight to that prediction if the patient was sedated and paralyzed, which would explain the low coma score.

Participants also wanted information about the baseline risk of mortality, the trend of mortality risk over time, and proper interpretation/use of the explanations (Table 12, Supporting information). The baseline risk of mortality was considered helpful in properly interpreting the predicted risk of mortality (e.g., a predicted risk of 50% is much more concerning when the baseline risk is 2%). Trends of mortality risk over time were also requested (mainly by nurses) to improve model utility, as it was noted that a change in risk was of more clinical interest than a single predicted risk (e.g., a patient that has had a high predicted risk for several days is of less concern than a patient with a lower predicted risk that has been recently increasing). Finally, although not specifically requested to be included in the explanation, several participants pointed out that training on interpretation of the prediction explanation would be vital to prevent improper use of the system (i.e., predictors are not suggestive of interventions to reduce mortality risk). Although useful in assessing credibility and utility, supporting information elements were not observed to have a specific influence on participant perceptions of the model.

Table 12. Insights on explanation design target questions

	Topic	Insight
What	Type, target, and level of explanation	*Providers sought "input", "output", "certainty", "why not", and "what if" explanations* "Do I get a confidence interval somewhere?" "The first question I get when I talk to the doctor is 'well, why did that happen?'" *Some desire for explanations of model processes and at the global-level* "What's the weight of the data that is available from the moment they did transfer them to the ICU and how does that carry into this predictive model?" "Is it possible to see what the machine learned about the relationship between age and raw numbers of vital signs?"
	Supporting information	*Raw feature values, raw time-series data, and contextual information aid interpretation* "Can I see the data for the blood pressure? Like a 57-point deviation in mean arterial pressure is quite substantial." "Everything bad that's happening with this patient seems to be contributed by the Coma Score…Now I see a diagnosis code of a brain tumor I'd give it much more weight." *Providers wanted to see baseline risk and trends of risk prediction over time* "This doesn't show like 'okay, the in-hospital mortality odds, 3 hrs ago was 1.6 and now it's actually coming down?" "I also think that an odds of death of 12% is still concerning—this child is 10 times more likely to die than the average child in our ICU." *Providers stressed importance of proper training on explanation interpretation* "One risk, I think, with this type of data presentation, is I are going to over-interpret the results…this is just showing you how the model worked, it doesn't necessarily mean the model is saying you should act on these specific [factors]"
How	Unit of explanation and organization	*Providers preferred feature groupings for the initial explanation display* "I really appreciate the groupings on this graph…I said well there's only an 8% risk of death but it's being driven largely by this neuro bucket, so what's going on there?" *Providers had mixed preferences on organization of predictors* "I like 2-2…I mean because that's how my brain thinks—I mean I break things down by that a lot of times in my head—physical assessments, labs, that sort of thing…" "Being able to see these are the top 5 increase mortality, these are the top 5 lower mortality...as opposed to the 2-2, where I can just see sort of graph fatigue"
	Dimensionality	*Providers wanted interactive linkage of data across plots and tables* "I also like that clicking on the graph directs you to the associated lab/physical assessment in the data table." "Is it possible that when you click on lactate results it could both bring up the raw data graph as well as the little components of the lactate table? *Interactive control over unit of explanation supported different information needs* "With the hover capability, if people wanted more, they could have that…I have a basic that everybody can get the basic data, but if you want to dig deeper it's there."
	Information representation	*Providers preferred risk to be presented as probabilities expressed as percentages* "I actually took out a calculator and based on the odds, I calculated the patient's percent probability of death." "For me, percent risk of mortality is going to be easier to interpret than the odds." *Providers preferred less statistical terminology for describing trend-based features* "It would be nice if I could just say, 'pulse ox is decreasing, creatinine is increasing'" *Visual cues play an important role in interpretation and design preferences* "It's because of the scale of the bars. The tighter bars just don't grab my attention." "I think with mock-up 1-3, my first initial thought is to look—you read left to right. So I would think the SpO monitor results would be the highest contributor." *Provider had mixed preferences on explanation display format* "I think 1-3 actually gives you the most visually, because you can see every element that's playing into that...I would suspect that there are other people who are going to look at 1-1 or 1-2 and it's going to be very obvious to them what they're looking at." "In my head, I'm thinking what about a pie chart?"

Participants exhibited strong preferences for the unit of explanation and risk representation design options, but had mixed preferences for other design options (Figure 17). In general, strong preferences on design options correlated with increased perceived importance of the design element (Figure 18). More specifically, all participants felt very strongly that risk be displayed in probabilities expressed as percentages (Figure 17; Figure 18; Table 12, Information representation). Expressing risk in terms of odds or as proportions seemed to increase the cognitive effort required to process the risk, thus negatively influencing the perceived usability of the system. All participants also expressed strong preferences for using feature groups (e.g., larger cognitive chunks) as the unit of explanation (Figure 17; Figure 18; Table 12, Information representation). Although the interactive option to view the individual features within each feature group was generally well-received (Table 12, Dimensionality), showing individual features in the initial explanation display appeared to require more cognitive effort to process and thus negatively influenced the perceived usability of the system. Some participants wanted even higher-level units of explanation than the feature groups, such as groups of related labs (e.g., electrolytes) or summary explanations akin to a human explanation (e.g., "This patient is at high risk for mortality because they are comatose, newly hypotensive, have been progressively hypoxic and have newly developed lactic acidosis").

The mixed preferences observed for the organization of explanation units (Figure 17; Table 12, Unit of explanation and organization) and explanation display format (Figure 17; Table 12, Information representation), and dimensionality design options (Figure 17) can be attributed to the different information needs and work environments of different clinical roles (Table 10, User cognition; Table 10, When/where). Nurses tended to prefer explanations that were static (i.e., had minimal information to process) and had explanation units organized by assessment (i.e.,

organized by controllability) (Figure 17). They found the tornado plots easier to understand than the force plots, and preferred simpler explanations (i.e., "The less you have to do the better"). These preferences all align with nurses' goal to extract actionable information (Table 10, Learning) within the cognitive and time constraints of their work environment (Table 10, Cognitive and time resources). On the other hand, physicians tended to prefer explanations that were modifiable (i.e., could provide more information upon request), which aligns with their desire to gain insights about a patient condition (i.e., using the model as an investigative tool). Physicians demonstrated no clear trend for explanation display format or explanation unit organization (Figure 17), which could simply be attributed to differences in how individual physicians prefer to process information. Explanations that fit the preferences of each clinical role contributed to positive perceptions of usability.

In addition to the design elements already noted to influence perceptions of usability, participants suggested some improvements to the design to increase the perceived usability. Specifically, participants utilized the plots of raw time-series data so frequently in interpreting explanation features that they requested multiple plots to view and compare data. They also requested that the points used to create features (e.g., min, max, most recent) be highlighted on the plots. Due to the importance of the baseline risk of mortality in interpreting the predicted risk, participants suggested that the it be more prominently displayed (e.g., above the predicted risk of mortality). To facilitate efficient data exploration, participants wanted data to be linked across interface elements (Table 12, Dimensionality) so that a selection of a data element (e.g., a laboratory test) in the explanation plot or raw feature table would auto-populate the time-series data plots with the selected data element. Finally, participants noted that comments on the vocabulary and visualization used suggested some possible influences on usability (Table 12,

Information representation). Specifically, participants noted a desire for simpler terminology for trend-based features (e.g., “Cr has increased since min value” rather than “Cr change from min >0.35”) and visualizations that obeyed standard information processing procedures. For example, the force plot was considered confusing because the increasing predictors were on the right and the largest increasing predictor was to the far right, violating standard ‘left to right’ reading procedures. Additionally, the scroll option on the explanation plot did not have a standard icon and some participants did not know that they could scroll to view additional predictors.

5.3 Final User-centered Explanation Design

Based on the insights in section 5.2, I proposed a final user-centered explanation design to be used in an evaluation study with target users. As nurses found more utility in risk trends than individual risk predictions with explanations, the proposed explanation design is intended to be used by physicians in assessing patient condition and priority. I leave exploration for the optimal presentation of model information to nurses for future work. Based on the insights from 5.2, I chose the following design options to maximize physician perceptions of the credibility, utility, and usability of the system:

1) Unit of explanation—feature groups for the initial display, where contributions of individual features for each time-series variable are aggregated by influence (e.g., whether they increase/decrease risk). Feature groups have an interactive hover-box option to view individual features comprising each group. In the interactive hover-box, trend-based features were summarized by whether they were an increasing or decreasing trend.

2) Organization of explanation units—default view of a single plot with explanation organized by decreasing magnitude of influence and increasing/decreasing risks grouped together by feature group. Interactive options were included to show groups of explanation units in the single plot by influence on risk or assessment type.
3) Risk representation—risks represented as probabilities expressed as percentages. The baseline risk was moved to a more prominent display location.
4) Explanation display format—tornado plot. While the mixed preferences of physicians for tornado and force plots would suggest that an interactive option to control explanation display format would also be beneficial, I opted not to include this option as some participants had found the force plot confusing to interpret.
5) Dimensionality—in addition to the interactive options to control the unit of explanation and the organization of explanation units mentioned above, the scroll option on the explanation plots was included to allow control over the number of explanation units viewed.

Per the insights in section 5.2, the explanation design included the table of raw feature values as well as two plots to display raw values of time-series data that highlighted the points used to generate features. As most relevant contextual information (e.g., diagnoses, interventions) requested by participants would be readily available in the EHR in which a system like this would be embedded, it was not included in the explanation design to reduce the amount of information presented on a single screen. For the evaluation study, relevant contextual information was provided in banner bar and a separate information tab (see Chapter 6). The final explanation design is shown in Figure 19.

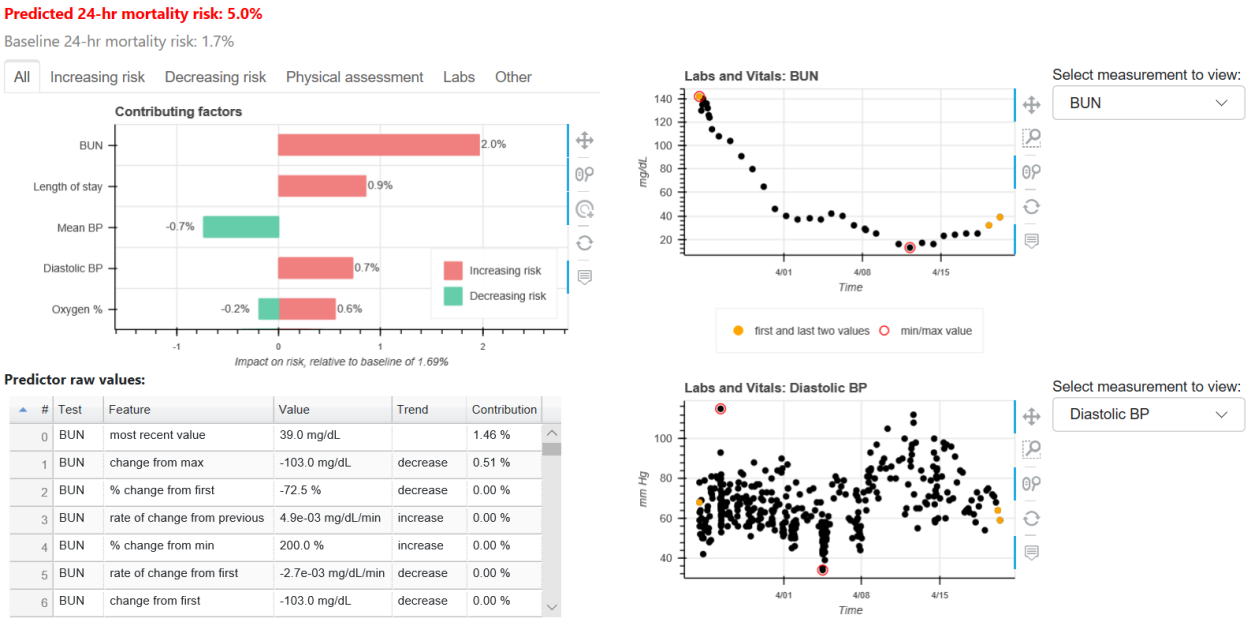


Figure 19. Final user-centered explanation design. The predicted risk and baseline risk are displayed at the top of the figure. The explanation plot (top left) uses feature groups as the explanation unit, but has hover-box capability to view individual features within each feature group. The plot includes interactive controls to view additional predictors and view sets of feature groups (e.g., view laboratory test feature groups). The raw feature table (bottom left) includes the description, value, and contribution to the risk for each individual feature. This table also includes the trend direction for trend-based features. The plots to display raw values for time-series features (right) highlight the points used to compute features and include interactive controls to zoom in on regions of data. These plots also have a hover funtionality that can be used to show the value and time of specific point. To facilitate data exploration, interactivity is linked across plots and tables (e.g., selecting a predictor on the explanation plot will highlight it in the raw feature table and load the appropriate laboratory test/vital sign in the time-series plot).

6.0 Evaluation

This chapter describes the evaluation of the user-centered explanation design for the pediatric ICU in-hospital mortality risk model that was described in Section 5.3. I conducted a laboratory study with healthcare providers to assess the impact of the user-centered explanation design on provider decision-making and perceptions of the model. Specifically, I examined the use of the model in assisting healthcare providers as they reviewed patient information in preparation for patient rounds. In this scenario, it was hypothesized that the prediction model with explanations could assist a healthcare provider in assessing patient condition and preparing to discuss a patient case with the rounding team. More specifically, when compared with the prediction model without explanations and having no prediction model, I hypothesized that the prediction model with the user-centered explanation design would improve healthcare provider:

1) accuracy in identifying patients who need to be seen urgently and in selecting relevant information to discuss with the rounding team
2) self-reported confidence in identifying patients who need to be seen urgently
3) efficiency in reviewing patient cases

Additionally, I hypothesized that relative to the prediction model without explanations, the prediction model with the user-centered explanation design would improve healthcare provider perceptions of the performance expectancy and effort expectancy of using the model in clinical practice.

6.1 Materials and Methods

This section describes the patient cases and participants, study design and tasks, data collection procedures, and data analyses for the laboratory study. The University of Pittsburgh IRB determined that this study was not considered human subjects research (STUDY19050287).

6.1.1 Participants and Patient Cases

A senior pediatric ICU attending selected six patient cases from the test dataset described in section 4.1.1 to be utilized in the laboratory study. Three cases included patients who needed to be seen urgently and three cases included patients who did not need to be seen urgently. The patient cases were selected to be clinically different such that independence of patient case could be assumed when performing data analyses.

A convenience sample of healthcare providers was recruited through professional connections of one of the dissertation committee members. Specifically, senior residents, fellows, and junior attending physicians (<1-year experience) specializing in critical care medicine were recruited to participate in the study. These groups were targeted because they were experienced in preparing for rounds on pediatric ICU patients.

6.1.2 Study Design and Tasks

I conducted a mixed-methods, within-subject evaluation study with three experimental conditions: 1) no access to information from the prediction model ("no model"), 2) access to model inputs and predictions from the prediction model ("prediction only"), and 3) access to

predictions and user-centered explanations from the prediction model ("explanation"). To conduct the evaluation study, I developed a local web-browser application that participants used to complete study tasks. The application is described in Section 6.1.3. Figure 20 provides an overview of the study design and tasks. Sessions were conducted with individual participants and lasted approximately 90 minutes. All sessions took place from September 2019-November 2019 and were conducted in a conference room near the participant's place of employment.

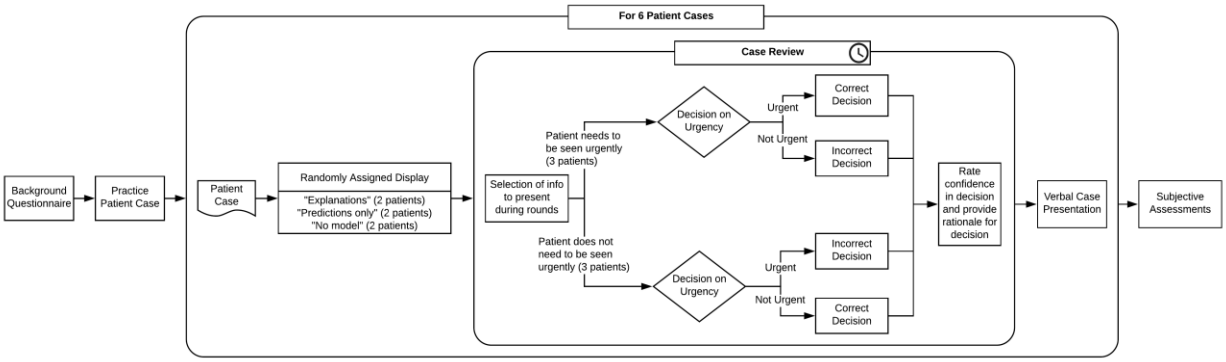


Figure 20. Overview of evaluation study design and tasks

Study sessions began with the participant watching a PowerPoint video presentation introducing them to the study objectives, the prediction model, and the study application. The slide deck from the video is provided in Appendix D. After the slideshow, participants were logged into the study application and asked to complete a short background questionnaire about their clinical experience. They were then provided access to a practice patient case that included mortality risk predictions accompanied by the user-centered explanation design and given time to familiarize themselves with the study application and ask any questions. Each participant was then asked to review each of the six patient cases, pretending as though they were preparing for patient rounds. For each case, participants were provided with a case vignette, diagnosis information, and access to laboratory test and vital sign results. Participants were also provided with one of three

different displays corresponding to the three experimental conditions: 1) "no model", which provided no additional information on the patient (Figure 21 in section 6.1.3), 2) "prediction only", which included a mortality risk prediction with model inputs but no explanation (Figure 22 in section 6.1.3), and 3) "explanation", which provided a mortality risk prediction accompanied by the user-centered explanation design described in section 5.3 (Figure 19 in section 5.3). Displays were randomly assigned to patient cases such that each participant reviewed two cases (one urgent, one non-urgent) with each one of the three displays. Participants reviewed each patient case only once, and patient cases were shown in a random order.

For each patient case, participants were asked to complete a questionnaire corresponding to the following tasks: 1) select the information they feel would influence changes in the plan of care for the patient (i.e., information they would want to present/discuss with the rounding team); 2) decide if the patient needs to be seen urgently by a member of the care team; 3) rate their confidence in the decision; and 4) provide a brief free-text rationale for the decision. After submitting the case response form participants were asked to verbally present their assessment of the patient as if they were presenting to the rounding team. After reviewing all six patient cases, participants were asked to complete subjective assessments of the "prediction only" and "explanation" displays, which included completing a subset of the UTAUT construct scale[38] item questionnaire to assess the performance expectancy and effort expectancy of each of the displays (see section 1.2 for a discussion on the relevance of these constructs). Participants were also offered the chance to provide unstructured feedback on each display during the subjective assessments.

Copies of the background questionnaire, patient case questionnaire, and UTAUT questionnaires are provided in Appendix E.

6.1.3 Study Application

To conduct the evaluation study, a local web-browser application was developed that would track participant progress on tasks, allow participants to interactively explore patient case information, and automate the data collection process. Each participant was assigned a unique login and password to access the study application. After logging in, participants were brought to a home page (Appendix D, slide 5) where they could track their progress on each of the study tasks. As shown in Figure 21, each patient case contained: 1) a banner bar providing basic demographic information about the patient; 2) a case information tab containing a case vignette, admitting diagnosis information, and access to plots to view laboratory test and vital sign data; and 3) a responses tab where they could complete the patient case questionnaire. This information was all that was provided for the "no model" display option.

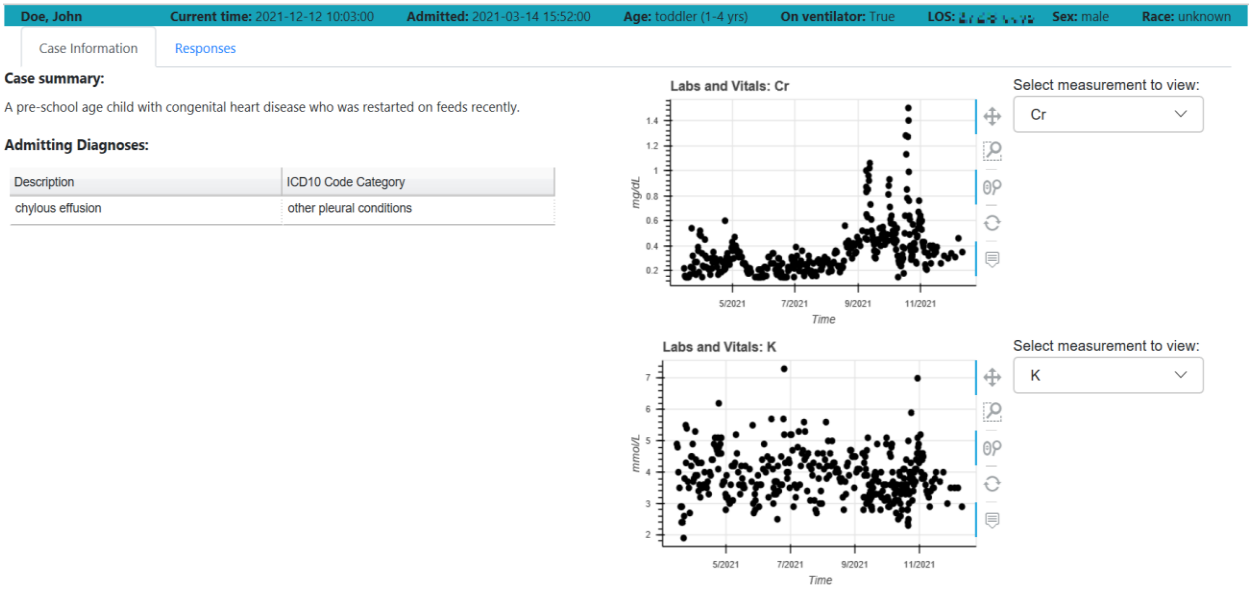


Figure 21. "No model" display that contains information available for every patient case

When participants had access to information from the pediatric ICU in-hospital mortality risk model, they were provided with an additional tab that contained information related to either the "predictions only" display, which is shown in Figure 22, or the "explanation" display, which is shown in Figure 19 in section 5.3. For each participant and patient case, the application recorded time-stamped interactions with interface elements (e.g., tabs, plots, tables).

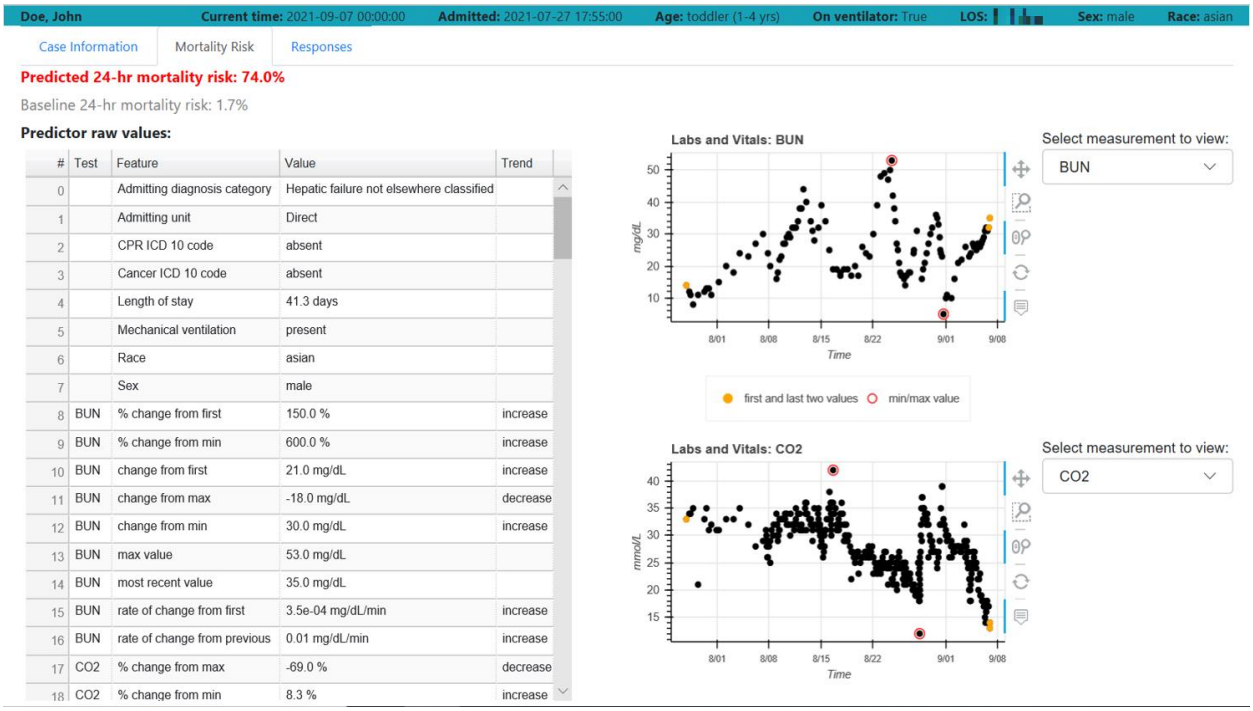


Figure 22. "Predictions only" display with additional tab containing mortality risk information

The application was developed in Python 3, utilizing the Python packages Flask version 1.0.2,[130] bokeh version 1.1.0,[127] shap version 0.28.5,[126] WTForms version 2.2.1,[131] and SQLAlchemy version 1.3.3.[132] Flask is a Python-based web framework that formed the backend of the application and was responsible for processing user input and producing the correct output (i.e., correctly routing user requests, processing input from forms). By default, Flask uses the Jinja2 templating engine, which facilitates dynamic HTML pages. The Flask extension for the popular Bootstrap front-end framework was used to style the HTML in a consistent manner. The bokeh

package was used to generate the interactive data displays used in the interface, i.e., the time-series plots for labs/vitals, the explanation plot, and the predictor tables. To collect responses from participants, the Flask extension for the WTForms Python library was used. The shap package was used to pre-generate patient explanations for each of the patient cases and all patient case data was stored within a local SQLite database. The Flask extension for the SQLAlchemy package was used to access the SQLite database to retrieve data, record responses, and track interface interactions.

6.1.4 Data Collection

Table 13 summarizes the data collected for each study task. It should be noted that the original scale items for the key UTAUT constructs in the subjective assessment task (performance expectancy, effort expectancy) were experimentally selected from the scale items of constructs from other models of technology acceptance and use (called root constructs). Only a few scale items were originally selected for each key construct and not all scale items were relevant to assess in the context of the proposed experiment (e.g., performance expectancy scale item of "If I use the system, I will increase my chances of getting a raise"). Therefore, for each key construct, I selected a set of scale items from each respective root construct that were relevant to assess in the context of the proposed experiment.

Table 13. Data collected for each study task

Study Task	Data Collected
Background Questionnaire	• Current clinical position (e.g., resident) • Length of time in current position (e.g., <1 year)
Patient Case Review	Data collected for each patient case: • Time-stamped interactions with application interface (e.g., tab selections, lab tests viewed) • List of information selected to discuss during rounds • Urgency decision accuracy (see Figure 20) • Urgency decision confidence, rated from 1—not confident at all to 5—extremely confident • Free-text rationale for urgency decision • Time (in seconds) to review patient case (excludes verbal case presentation, see Figure 20) • Audio-recording of verbal patient case presentation • Moderator notes on interesting comments or behavior during case review
Subjective Assessments	Data collected for "prediction only" and "explanation" displays: • Selected UTAUT Root Construct Scale Items for Performance Expectancy[38] (Likert scale agreement): 1. Using the system would enable me to accomplish tasks more quickly. 2. Using the system would make it easier to do my job. 3. Using the system would increase my productivity. 4. I would find the system useful in my job. • Selected UTAUT Root Construct Scale Items for Effort Expectancy[38] (Likert scale agreement): 1. My interaction with the system would be clear and understandable. 2. I would find the system easy to use. 3. It would be easy for me to become skillful at using the system. • Free-text feedback on the display (optional)

6.1.5 Data Analysis

Audio recordings of all verbal case presentations were transcribed verbatim and compiled with urgency decision rationales and moderator notes for each case. Answers to background questionnaires were summarized in a contingency table. Based on the background questionnaire responses, two levels of clinical experience (residents and fellow/attendings) were defined for use in analyses. Primary outcomes of interest included the impact of the user-centered explanation display on decision accuracy, decision confidence, case review efficiency, and provider perceptions of the pediatric ICU in-hospital mortality risk model. Analyses for each outcome are summarized in Table 14 and described in the next few sections. P-values of <0.05 were considered

significant for all statistical analyses, which were carried out using Stata version 15.[133] Plots were generated using the Python packages seaborn version 0.9.0[134] and matplotlib version 3.0.3.[135]

Table 14. Summary of analyses examining the impact of the user-centered explanation display on outcomes

Outcome	Display Comparison Groups	Metrics	Analytic approach
Decision accuracy	"No model" "Prediction only"	Urgency decision accuracy	Proportion of correct decisions with 95% CI Logistic mixed effect analysis
		Precision and recall in selecting relevant information	Visual review of violin plots
		Mentions of predictive model in rationales, transcripts, or notes	Qualitative review to assist in interpretation of quantitative results
Decision confidence	"No model" "Prediction only"	Urgency decision confidence	Visual review of stacked bar charts Ordinal logistic mixed effects analysis
		Mentions of predictive model in rationales, transcripts, or notes	Qualitative review to assist in interpretation of quantitative results
Case review efficiency	"No model" "Prediction only"	Time to review patient case	Descriptive statistics Log-linear mixed effects analysis
		Number of unique items viewed (computed from interactions data)	Descriptive statistics Poisson mixed effects analysis
		Total number of items viewed (computed from interactions data)	Descriptive statistics Negative binomial mixed effects analysis
Provider perceptions	"Prediction only"	UTAUT questionnaire responses	Visual review of stacked bar charts
		Free-text feedback on displays and moderator notes	Qualitative review for insights about participant perceptions of predictive model

Analysis of decision accuracy

Decision accuracy included participant accuracy in urgency decisions (i.e., identifying patients who need to be seen urgently) as well as selecting relevant information to discuss with the rounding team. To evaluate urgency decision accuracy, the proportion of correct decisions with 95% CIs for each of the three displays were calculated and a logistic mixed effects analysis of the relationship between urgency decision accuracy and display was performed. Display, case urgency (urgent, non-urgent), and participant experience (resident, attending/fellow) were included as fixed effects in the model (no interaction terms), and an intercept for participant was included as a random effect in the model. To assess accuracy in selecting relevant information, participant

precision and recall in selecting 'relevant' items were calculated, where information items selected by a senior pediatric ICU attending using the "explanations" display served as the gold standard. Precision and recall scores for each display were visualized using violin plots. Decision urgency rationales, case presentation transcripts, and moderator notes were reviewed for mentions of the predictive model tool and to assist in interpretation of the results.

Analysis of decision confidence

To assess the relationship between the display shown and participant-reported confidence in their urgency decision, confidence ratings for each of the displays were visualized in a stacked bar chart and an ordinal logistic mixed effects analysis was performed. Display, case urgency (urgent, non-urgent), and participant experience (resident, attending/fellow) were included as fixed effects in the model (no interaction terms), and an intercept for participant was included as a random effect in the model. Decision urgency rationales, case presentation transcripts, and moderator notes were reviewed for mentions of the predictive model tool and to assist in interpretation of the results.

Analysis of case review efficiency

Case review efficiency consisted of the time it took participants to review each patient case and the amount of information being viewed, which was measured by the number of items (e.g., lab test, vital sign) viewed during the case. Descriptive statistics were used to summarize the case review time, number of unique items viewed, and the total number of items viewed. To assess the relationship between the display shown and case review time, a log-linear mixed effects analysis was performed after it was determined that case review time followed a log-normal distribution.

To assess the relationship between the display shown and the number of unique items viewed, a Poisson mixed effects analysis was performed. To assess the relationship between the display shown and the total number of items viewed, a negative binomial mixed effects analysis was performed after it was determined that the distribution of the total number of items was over-dispersed (mean=33.0; variance=206.3). For all three models, display, case urgency (urgent, non-urgent), participant experience (resident, attending/fellow), and case order (i.e., the order in which the case was seen by a participant) were included as fixed effects (no interaction terms) and an intercept for participant was included as a random effect.

Analysis of provider perceptions

Responses to the UTAUT scale items for the "explanation" and "prediction only" displays were visualized and compared using stacked bar charts. Free-text feedback on displays and moderator notes were qualitatively reviewed to assist in the interpretation of the UTAUT questionnaire responses and to identify additional insights about participant perceptions of the pediatric ICU in-hospital mortality risk model and the displays.

6.2 Results

A total of 15 participants were recruited for this study. Responses to the background questionnaire on clinical experience are summarized in Table 15. As per the study design, each participant reviewed and provided responses for 6 patient cases. Due to a technical error, one participant failed to successfully complete one of their assigned cases. Thus, there were a total of 89 participant responses for the patient cases. The breakdown of case responses by display and

case urgency is shown in Table 16. In 6.2.1-6.2.3, I describe the results from the analyses on decision accuracy and confidence, case review efficiency, and provider perceptions of the model, respectively.

Table 15. Summary of participant clincial experience

	Time in current position			
Position	<1 year	1 to <2 years	2 to <3 years	**Total**
Attending	1	0	0	1
Fellow	1	5	1	7
Resident	0	2	5	7
				15

Table 16. Participant responses by case urgency and display

	Case Urgency		
Display	Non-urgent	Urgent	**Total**
No model	14	15	**29**
Prediction only	15	15	**30**
Explanation	15	15	**30**
Total	**44**	**45**	**89**

6.2.1 Decision Accuracy and Confidence

As shown in Table 17, the proportion of correct decision responses was highest with the "explanation" display; however, all proportions had substantially overlapping 95% CIs, which makes it challenging to comment on the significance of this effect. The results of the logistic mixed effects analysis (Table 18) detected no significant effect of display, case urgency, or participant experience on decision accuracy. As seen in Figure 23, neither the precision nor recall scores revealed discernable differences in provider accuracy in selecting relevant patient information.

Table 17. Proportion of correct decisions for each display

Display	Proportion of correct decisions	95% CI
No model	0.69	[0.49 – 0.85]
Prediction only	0.73	[0.54 – 0.88]
Explanation	0.87	[0.69 – 0.96]

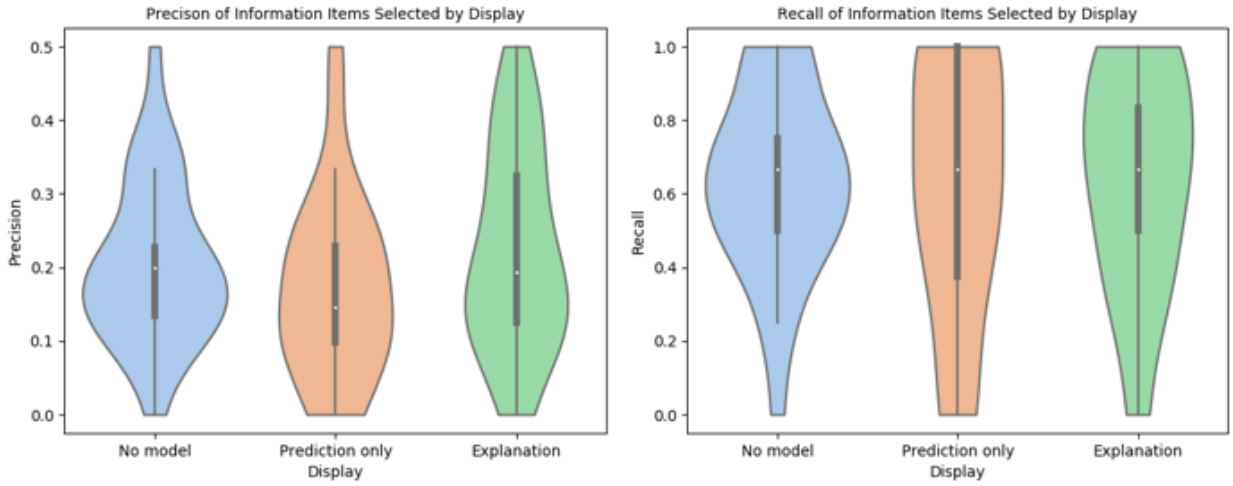


Figure 23. Participant accuracy in selecting relevant information. The plots show the distributions of precision (left) and recall (right) scores for each of the three displays.

As seen in Figure 24, participants seemed confident in their urgency decisions regardless of display, with all providers rating their confidence as a 3 or higher for all decisions. While Figure 24 suggests that providers might be more confident in their decisions when they had access to a mortality risk prediction ("predictions only" and "explanation" displays), the ordinal logistic mixed effects analysis (Table 18) detected no significant effect of display, case urgency, or participant experience on decision confidence. The lack of display impact on decision accuracy and confidence is further supported by the fact that decision rationales and case presentations contained relatively few mentions of the pediatric ICU in-hospital mortality risk model (six total mentions by four different participants). When mentioned, participants were either questioning the

validity of a model prediction with the "prediction only" display (e.g., "mortality risk is lower, but it looks like a lot of things are kind of trending in the 'not right' direction") or using the mortality risk prediction from either the "prediction only" or "explanation" display as part of their justification for the urgency decision ("a child who is under-supported with a high risk of mortality").

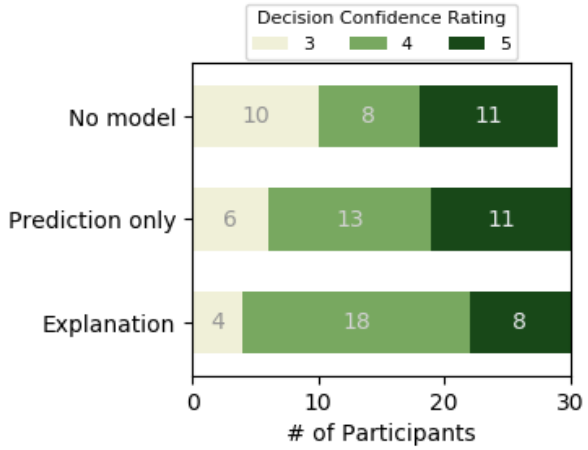


Figure 24. Provider self-reported confidence in urgency decisions for each display

Table 18. Summary of the analyses of display effect on decision accuracy and decision confidence

		Logistic mixed effects analysis of display effect on decision accuracy				Ordinal logistic mixed effects analysis of display effect on decision confidence			
Random effect		**Variance**				**Variance**			
Participant (intercept)		4.71e-34				1.50			
Fixed Effects		**Odds Ratio**	**Std. Error**	**p-value**	**95% CI**	**Odds Ratio**	**Std. Error**	**p-value**	**95% CI**
Display	No model	Reference				Reference			
	Prediction only	1.23	0.71	0.71	[0.40, 3.83]	1.56	0.84	0.41	[0.54, 4.46]
	Explanation	2.93	1.97	0.11	[0.79, 10.93]	1.26	0.67	0.66	[0.45, 3.54]
Urgency	Not urgent	Reference				Reference			
	Urgent	1.18	0.60	0.74	[0.44, 3.20]	1.81	0.78	0.17	[0.77, 4.22]
Experience	Attending/fellow	Reference				Reference			
	Resident	1.18	0.60	0.75	[0.43, 3.21]	0.50	0.39	0.38	[0.11, 2.29]

6.2.2 Case Review Efficiency

As seen in Table 19, the "explanation" display had the longest average case review time and lowest average number of items viewed (both unique and total), but the analyses demonstrated that there was no significant effect of the display shown on case review time, the number of unique items viewed, or the total number of items viewed (Table 20 and Table 21). The analyses did reveal a significant effect of case urgency on case review time and the unique number of items viewed (Table 20 and Table 21). More specifically, when controlling for other factors, participants spent a longer time reviewing a case and viewed more unique items for urgent cases when compared to non-urgent cases. The analyses also revealed a significant effect of the order in which the case was seen by a participant on case review time and the total number of items viewed (Table 20 and Table 21). More specifically, when controlling for other factors, participants spent less time per case and viewed less total items per case as they progressed through the set of six patient cases, i.e., participants became more efficient in their case review as they went through the study tasks.

Table 19. Mean and variance of case review efficiency measures for each display

	Efficiency Measure Mean (variance)		
Display	**Case review time in minutes**	**# of unique items viewed**	**Total # of items viewed**
No model	8.0 (20.5)	22.2 (9.5)	34.5 (324.1)
Prediction only	7.7 (17.8)	22.0 (7.9)	33.3 (208.5)
Explanation	8.8 (27.4)	21.5 (31.8)	31.3 (99.3)

Table 20. Summary of the analysis of display effect on case review time

		Log-linear mixed effects analysis of display effect on case review time			
Random effect		**Variance**			
Participant (intercept)		0.15			
Fixed Effects		**Coefficient**	**Std. Error**	**p-value**	**95% CI**
Display	No model	Reference			
	Prediction only	0.08	0.08	0.36	[-0.09, 0.24]
	Explanation	0.04	0.08	0.36	[-0.11, 0.19]
Urgency	Not urgent	Reference			
	Urgent	0.20	0.06	**0.002**	[0.07, 0.33]
Experience	Attending/fellow	Reference			
	Resident	-0.01	0.21	0.95	[-0.43, 0.40]
Case Order	1	Reference			
	2	-0.28	0.09	**0.002**	[-0.46, -0.11]
	3	-0.47	0.10	**0.000**	[-0.66, -0.28]
	4	-0.63	0.11	**0.000**	[-0.86, -0.42]
	5	-0.72	0.11	**0.000**	[-0.95, -0.50]
	6	-0.81	0.12	**0.000**	[-1.05, -0.56]

Table 21. Summary of the analyses of display effect on unique and total number of items viewed

		Poisson mixed effects analysis of display effect on unique number of items viewed				Negative binomial mixed effects analysis on total number of items viewed			
Random effect		**Variance**				**Variance**			
Participant (intercept)		0.003				0.02			
Fixed Effects		**Rate Ratio**	**Std. Error**	**p-value**	**95% CI**	**Rate Ratio**	**Std. Error**	**p-value**	**95% CI**
Display	No model	Reference				Reference			
	Prediction only	0.99	0.06	0.89	[0.89, 1.11]	1.04	0.09	0.62	[0.88, 1.23]
	Explanation	0.97	0.05	0.56	[0.87, 1.08]	0.92	0.08	0.34	[0.79, 1.09]
Urgency	Not urgent	Reference				Reference			
	Urgent	1.11	0.05	**0.02**	[1.02, 1.22]	0.99	0.07	0.93	[0.87, 1.14]
Experience	Attending/fellow	Reference				Reference			
	Resident	0.93	0.05	0.17	[0.83, 1.03]	0.85	0.08	0.10	[0.71, 1.03]
Case Order	1	Reference				Reference			
	2	1.03	0.08	0.66	[0.89, 1.21]	0.79	0.09	**0.05**	[0.63, 1.00]
	3	0.94	0.08	0.46	[0.80, 1.10]	0.68	0.08	**0.001**	[0.54, 0.86]
	4	1.01	0.08	0.92	[0.86, 1.18]	0.72	0.08	**0.005**	[0.57, 0.91]
	5	0.99	0.08	0.92	[0.85, 1.16]	0.68	0.08	**0.001**	[0.54, 0.86]
	6	0.98	0.08	0.84	[0.84, 1.16]	0.64	0.08	**0.000**	[0.50, 0.81]

6.2.3 Perceptions of Prediction Tool

Figure 25 summarizes participant responses to the UTAUT questionnaire for the "prediction only" and "explanation" displays. In general, participants had positive perceptions of the performance expectancy of a system utilizing either the "prediction only" or "explanation" display (Figure 25, statements 1-4), but the "explanation" display improved participant perceptions of performance expectancy relative to the "prediction only" display. More specifically, a majority of participants thought that a system utilizing the "explanation" display would be useful in their job (93%), make it easier to do their job (73%), and increase their productivity (60%) (Figure 25, statements 1-3). In contrast, less than half of participants reported the same thoughts about the "prediction only" display (33%, 33%, and 46%, respectively). Several participants mentioned that a system without explanations would not be useful to them, specifically because they could not rationalize the prediction and identify why the patient might be at higher or lower risk. Participant comments indicated that the positive perceptions of the performance expectancy of the "prediction only" display were related to the benefit of having a mortality risk score to help in patient prioritization. One participant specifically commented on their positive ratings for the "prediction only" display:

> *"Both systems are already markedly better than our current electronic medical record (thus I marked all as strongly agree)."*
> *—2nd year fellow*

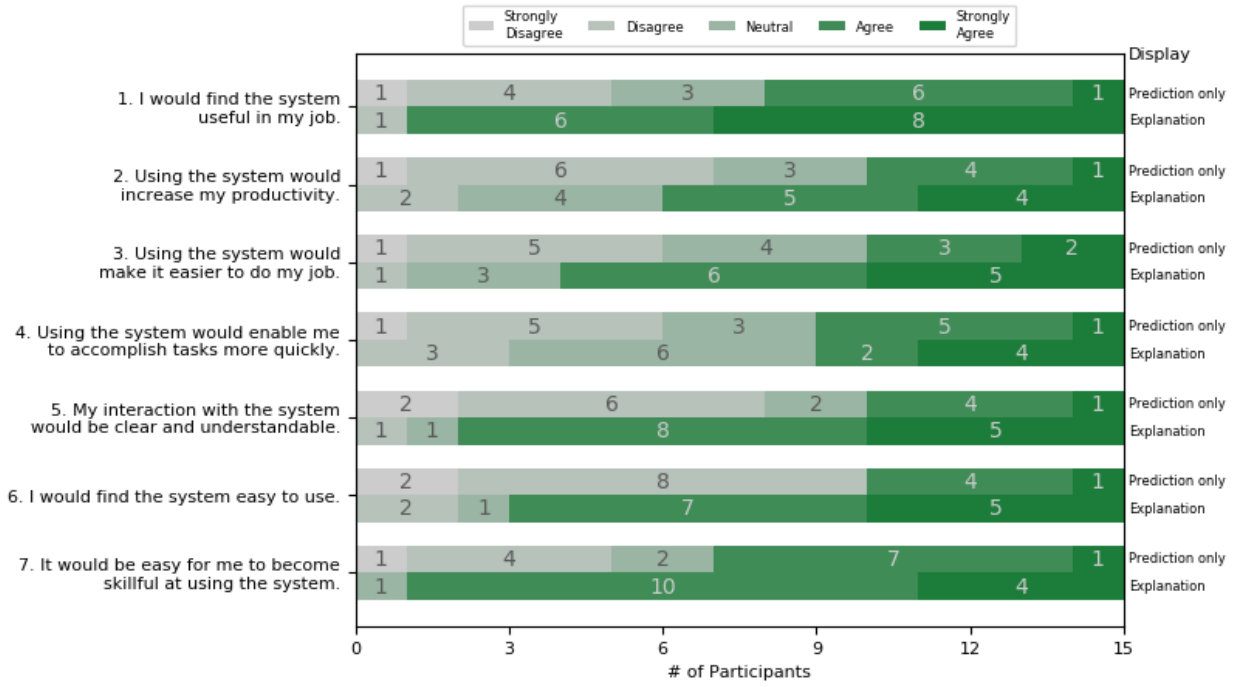


Figure 25. Participant responses to UTAUT questionnaire for "prediction only" and "explanation" displays. Statements 1-4 assess perceptions of performance expectancy and statements 5-7 assess perceptions of effort expectancy.

Despite the generally positive views of performance expectancy, only 40% of participants reported that a system utilizing either display would enable them to accomplish tasks more quickly (Figure 25, statement 4). Comments from participants revealed that this perception may have been partially influenced by having to adjust to unfamiliar data displays for laboratory test and vital sign data. In particular, a few participants commented that it took them longer to find and review raw data than it would have taken them in the EHR (e.g., having to look up each individual component of a basic metabolic panel), which could have negatively impacted participant perceptions of the system's ability to help them accomplish tasks more quickly. Moreover, although some participants thought the system would help them more efficiently assess a patient's condition, many participants seemed not to trust the system as a guide, viewing it instead as a tool to confirm

their own assessments. This view might explain why participants felt the system would not aid them in accomplishing tasks more quickly. One participant succinctly summarized this viewpoint:

> *"The explanations certainly help me dive under the black-box nature of the model without explanations, but I don't think it would dramatically improve my productivity. At this point, I would still want to evaluate each feature using my standard process, then look at the model to see if I missed anything, rather than using the model as a hypothesis generator. There were features that I was much more concerned about than the model, and vice versa, that builds inherent distrust."*
> *—2nd year fellow*

Overall, the "explanation" display greatly improved participant perceptions of effort expectancy relative to the "prediction only" display (Figure 25, statements 5-7). While some participants reported positive perceptions of the effort expectancy of the "prediction only" display, a large number of participants commented that the explanations greatly improved their ability to make sense of the risk prediction. Several participants commented that the information in the "prediction only" display was overwhelming and not helpful in understanding the prediction. One participant specifically noted his frustration with the "prediction only" display when he disagreed with a high risk prediction, commenting that the provided information did not help him understand why the model was showing an increased risk. Many participants expressed preference for the "explanation" display, stating that the explanation facilitated model interpretation and comparison with their own clinical judgment. One participant effectively summarized these ideas:

> *"[The 'prediction only' display made it] nearly impossible to tell what the key factors were—I'm simply drowning in the computer's output without a framework to make sense of it…[It's] very helpful to graphically demonstrate key drivers that the machine found important—it allowed me to integrate the machine's understanding with my own clinical intuition and knowledge of the patient's overall context, which I fear may be missed by a machine-learning model at times."*
> *—2nd year resident*

In addition to clarifying perceptions about the model and displays, participant comments and feedback identified a few possible design improvements for the "explanation" display. First, as already mentioned above, participant comments suggested that it may be beneficial to present laboratory test and vital sign data in the same format it is presented in the EHR (e.g., a fishbone diagram for basic metabolic panel data). Additionally, for the laboratory test and vital sign data views, participants requested: 1) the ability to plot multiple tests on the same plot, 2) 'quick buttons' to zoom in on relevant time ranges of data (e.g., last 12, 24, or 48 hours); 3) drop-down boxes that allow selection by groups of related results (e.g., electrolytes); 4) the ability to select points to be highlighted (e.g., non-selected points are "greyed out"); and 5) a table or list of current values next to the plot. In addition to feedback on the laboratory and vital sign test views, a few participants suggested improvements to interface interactions (e.g., using arrow keys to navigate a dropdown list) and several participants requested additional information about the patient to assist in their interpretation of the data (e.g., laboratory tests, ventilator settings, interventions).

7.0 Discussion

In this dissertation, I aimed to utilize clinician perspectives to inform the design of explanations for ML-based prediction tools to improve the adoption of these tools in clinical practice. Toward that goal, I developed a new theoretical framework of explanation design for ML models and used the framework in conjunction with healthcare provider feedback to inform the design of a user-centered explanation for predictions from a pediatric ICU in-hospital mortality risk model. The user-centered explanation design was a model-agnostic, instance-level explanation of feature influence generated using the publicly available SHAP algorithm.[124,125] I hypothesized that the predictive model with the user-centered explanation design would improve provider perceptions of utilizing the predictive model in practice, which would result in the predictive model improving provider accuracy, confidence, and efficiency in making decisions during preparations for patient rounds relative to having no model and having a model that does not provide explanations. While the results of the studies demonstrated that the user-centered explanation design improved provider perceptions of utilizing the predictive model in practice, no significant effect of the user-centered explanation design on decision-making accuracy, confidence, or efficiency was observed.

Overall, the results of the studies revealed that critical care providers had positive perceptions of the pediatric ICU in-hospital mortality risk model and the user-centered explanation design. In the qualitative inquiry study, providers found the mock-ups of the SHAP explanations useful in assessing the credibility and utility of a prediction from the model, (i.e., comparing the influential risk factors to domain knowledge to determine if the prediction seemed reasonable and clinically relevant). In the evaluation study, the user-centered design of the SHAP explanations

greatly improved provider perceptions of the performance expectancy and effort expectancy of using the pediatric ICU in-hospital mortality risk model in practice. These findings suggest that model-agnostic, instance-level, explanation approaches based on feature influence methods are a viable approach to explaining model predictions in a way that is both comprehensible and useful to healthcare providers. Although other studies have utilized these approaches to explain predictive models in healthcare,[10,16,75] to the best of my knowledge this is the first study to verify that these explanations would be positively received by healthcare providers. Provider acceptance of these explanations could help overcome the model interpretability barrier to utilizing ML models in practical applications in medicine. It should be noted that concerns about model interpretability as a barrier to adoption generally come after an ML model has been demonstrated to have acceptable performance, generalizability, and/or reproducibility (i.e., once a "good" model has been developed). However, model interpretability can assist model developers in ensuring these criteria are met. These criteria were considered outside the scope of the conducted studies, but are important to note in discussions of the value of interpretability when utilizing predictive models in healthcare.

Although providers indicated that they would accept and use the pediatric ICU in-hospital mortality risk model, the evaluation study revealed no significant effects of the model with the user-centered explanation display on decision-making accuracy, confidence, or efficiency, which was unexpected. However, the study was likely under-powered to detect effects in these outcomes, unless the effect size was very large. For example, let's consider the comparison of decision accuracy in only the "no model" and "explanation" display groups, where there would ideally be a total of 60 observations (30 in each group). Assuming the true proportion of correct responses in the "no model" display group was 0.68 (from Table 17) and assuming a total 60 observations (30

per group), for a chi-square test to detect a statistically significant effect at an alpha of 0.05 and power of 80%, the proportion of correct responses in the "explanation" display group would have had to have been 0.95 or higher. Similarly, treating the other outcomes as continuous and comparing only the "no model" and "explanation" display groups, a paired t-test with a total sample size of 60 (30 in each group), an alpha of 0.05, and a power of 80% would be able to detect an effect size of 0.74. Assuming the lowest standard deviation for each efficiency measure in Table 19 and a standard deviation of 0.5 for decision confidence, this would be a minimal difference of ~3 minutes in case review time, ~2 items in the number of unique items viewed, ~7 items in the number of total items viewed, and 0.37 points in decision confidence. In light of these analyses, it is less surprising that the study did not detect any significant effect of the user-centered explanation display on decision-making outcomes.

The analyses did detect significant effects of case urgency and the order in which a case was viewed on the decision efficiency metrics. Specifically, providers spent significantly more time reviewing a case and viewed significantly more unique items for urgent cases than for non-urgent cases. This was not surprising to find, as the urgent cases represented more medically complex patients and would likely require more review time by providers. It was surprising to find that providers spent significantly less time and reviewed significantly fewer total items for cases that were viewed later in the study session. This could have been because providers were still spending time to familiarize themselves with system after the practice patient case. Alternatively, the study sessions may have been too short to allow adequate time for providers to review all of the patient cases, which may have caused them to rush through the material as they neared the end of the session. This suggests that there was a learning curve required to use the system that was not adequately accounted for, which may have obscured the effect the shown display had on

measures of decision efficiency. Two possible ways in which the learning curve effect could have been better controlled include: 1) running pilot studies to estimate the time it would take participants to complete each study task and planning sessions of adequate length, and 2) developing a way to assess participants' comfort level with the system and requiring that they reach a predefined level of comfort during the practice patient case activity.

In contrast to the quantitative analyses, the subjective assessments of the performance expectancy of the systems suggested that providers would find the predictive model with explanations beneficial in performing their jobs, indicating that the tool would provide some benefit to decision-making. Two main viewpoints emerged about the benefit of the tool in decision-making. Some providers saw the tool as a confirmatory tool, i.e., a tool to confirm thought processes and check for things that they might have missed during their initial assessment. This viewpoint agrees with findings from Jeffery et al.,[136] who found that nurses mainly viewed probability-based CDSSs as a tool to confirm their thoughts about a patient, and Hallen et al.,[95] who found that providers perceived clinical prediction models as tools to improve their prognostic confidence. This would suggest that providers view predictive models as confirmatory tools to increase decision confidence, rather than as informatory tools to guide decisions. If used as a confirmatory tool, the system would have minimal impact on provider decision-making processes, which would partially explain the lack of observed effect on decision-making accuracy and efficiency and likely explains participants' generally low expectations that the tool would improve their ability to accomplish tasks quickly.

The second viewpoint about the benefit of the tool relates to decision efficiency. In particular, some participants viewed the predictive model with explanations as a useful tool to guide their assessment of patients and to prioritize patients. More specifically, by highlighting

patients and information of concern, providers thought the tool could help them be more efficient in prioritizing patients and reviewing patient information. This viewpoint likely explains the few participants who perceived the tool as something that would improve their ability to accomplish tasks quickly. Interestingly, the same number of participants reported that the "prediction only" and "explanation" display would improve their ability to accomplish tasks quickly. This suggests that the perceived improvement in task efficiency may stem from having a mortality risk prediction to help them prioritize patients, regardless of whether the prediction is accompanied by an explanation. This view can explain the positive perceptions of the performance expectancy observed for the "prediction only" display as the system was still providing new information to the participants, even though the system was less favorably perceived than the "explanation" display. This viewpoint contradicts past experiences in which high performing models have gone unused due to lack of interpretability, which suggests that user and environmental characteristics likely influence the types of predictive model systems that may be accepted and used.

Despite some participants' favorable perception of the performance expectancy of the "prediction only" display system, provider feedback on the system demonstrated that the "explanation" display system had higher performance expectancy and a much lower effort expectancy. Specifically, providers found that the explanations simplified interpretation of the risk prediction from the model and integration of the model information with their clinical knowledge. Several participants mentioned that they would not use the system without the explanations, particularly because they could not investigate predictions that surprised them. Although the benefit of providing explanations could not be demonstrated quantitatively, the subjective assessments and provider feedback suggest that it is still valuable to provide the explanations for the predictive risk model.

The discrepancy between the results of the subjective assessments of provider perceptions of the model and the quantitative analyses of the impact on decision-making raise several thoughts about the study and how to measure the potential value of a predictive model in clinical practice. First, it suggests that the study possibly targeted the wrong outcome metrics to assess the impact of the predictive model with explanations on provider decision-making. The provider view that the system would be useful in assessing and prioritizing patients suggests that perhaps the value of the tool is in comparing urgency levels of patients rather than assessing the urgency level of an individual patient, (i.e., making decisions about groups of patients rather than individual patients). If this is the case, the study missed an opportunity to examine the impact of the predictive model with explanations on an important decision-making process. Conducting a more thorough investigation of how the tool would be used in practice could have helped identify the appropriate decision-making process and outcomes to assess in the evaluation study.

Alternatively, it is possible that the study examined the right decision-making process, but the predictive model with explanations did not provide information that would directly influence provider decisions. This is supported by the provider view that the system was useful as a confirmatory tool and could increase decision confidence, but would not directly influence a decision. Shah et al.[5] suggest that many predictive models are not deployed into practice because they do not provide information that influence decisions. This would suggest that a predictive model tool would need to demonstrate a clear benefit on decision-making performance to be accepted in practice, but the results of the evaluation study showed that providers would still use and accept a tool even if it did not directly inform decision-making. Dekker et al.[137] mention that access to accurate risk predictions seems to have an unpredictable effect on provider decisions and thus claim that the true value of a predictive model tool cannot be known without running impact

studies to assess the effect on patient outcomes. This raises several interesting questions about how to evaluate predictive model tools for use in clinical practice. Specifically, what are the appropriate metrics and assessments for demonstrating the value of a predictive model tool? What evidence of the value of a predictive model tool should be required before it is deployed into clinical practice? If a predictive model does not demonstrate improved performance in some measureable way, do subjective assessments of provider satisfaction with the tool provide enough evidence of its value? These questions warrant further consideration as more predictive model tools make their way into clinical practice.

More generally, this work contributes to knowledge about the effective communication of predictive model risk information to healthcare providers. In both the qualitative inquiry and evaluation studies, it was found that providers liked the ability to visually assess which risk factors were contributing most to an individual's predicted risk. This finding provides evidence to support claims in the literature that visualizations of risk information for individuals can improve healthcare provider interpretation and acceptance of predictive models.[96,113] Additionally, results from the study revealed that providing the appropriate contextual information was vital to provider interpretation of risk. In particular, access to raw patient data (e.g., laboratory values, vital signs, interventions) played a significant role in provider ability to assess the clinical credibility and utility of predictions and explanations. This finding is consistent with results from studies by Wang et al.[40] and Jeffery et al.,[136] both of whom also found that providers utilized raw patient data when working with probability-based decision support systems to verify information from the system and integrate it with their clinical knowledge. In addition to raw patient data, several nurses in the qualitative inquiry study noted the importance of having a baseline risk and risk trends to assess the clinical relevance of a risk prediction. More specifically, a change in risk from a patient-

specific or population baseline was deemed more clinically relevant than a single risk prediction. This finding is consistent with results from Jeffrey et al.,[136] who also found that nurses wanted to see risk trends when using probability-based CDSS. While most research has focused on developing high-performing risk prediction models, these findings suggest the need for more research on how the manner in which risk information is communicated affects provider interpretation and use of the information in clinical practice.

The studies also revealed that interactive explanations of risk were beneficial to supporting different user information needs and preventing information overload by allowing users to ask for additional details when desired. Allowing users to ask for more information from the system improved participant perceptions of the system, which provides some support for claims that effective explanations for AI systems would mimic human explanation and occur as part of a social interaction or conversation.[27] Providing interactive explanations would also facilitate inclusion of multiple explanation types into a single explanation display, such as incorporating the "what if" type explanations that some participants requested during the qualitative inquiry study. While this type of explanation was not incorporated in the final design, an interactive explanation could support the inclusion of the additional type (e.g., adding interactive features to the SHAP explanation that allow users to change model inputs to see the change in predicted risk). The need for integrating multiple explanation types into a single explanation design has also been mentioned by Wang et al.,[40] who found that providers utilized a variety of different explanations to support various reasoning processes when diagnosing patients. The proposed theoretical framework in this work could support further exploration of how to design combinations of explanations that effectively support healthcare provider explanation needs in various tasks.

While this work advocates the need for user-centered explanation designs for predictive models in healthcare, one could argue that there may be scenarios in which explanations are not required at all. For example, if a model was hypothetically able to achieve perfect performance on a prediction task, explanations might be considered unnecessary. Alternatively, explanations might be considered unnecessary when it is obvious whether the model correctly predicts the outcome (e.g., image classifications that can be verified by visual inspection). In either of these scenarios, one could argue that explanations might still be needed to instill user trust in the model (i.e., a user may want to ensure that predictions can be justified based on domain knowledge). However, Elish[24] contradict this argument by pointing out that trust in a model can also be built by involving stakeholders throughout the model development process. Even if global explanations of a model are considered unnecessary, instance-level explanations could still provide valuable information to a user that enables them to act on a specific prediction. For example, consider a model that perfectly predicts a patient's risk of mortality. There are several different ways in which a patient may die and the ability of a provider to intervene will depend on the reason a patient is predicted to die. Thus, instance-level explanations may still prove valuable for the perfectly performing model. Although it is likely that some form of explanation will be required for most predictive models in healthcare, the need for explanations will be context-dependent and should be discussed in the early stages of model development.

7.1 Limitations and Future Work

The main limitation of this work was in the evaluation study design, specifically the small sample size and the insufficient study session length, which likely negatively impacted the ability

to detect any significant effect of the explanation design on decision-making outcomes. Additionally, it appears that the overall system design did not adequately represent the context in which the predictive model might be used in clinical practice. More specifically, the predictive model system was not presented as a tool integrated into the existing EHR, which is how the model would likely be presented to providers when deployed into practice. As noted by participant comments and feedback on the system, this presented issues due to lack of access to certain patient data (e.g., interventions) and unfamiliar data presentations (e.g., laboratory and vital sign plots). The lack of contextual information and unfamiliar data displays likely negatively impacted measures of the decision-making metrics, specifically decision efficiency. It may have also negatively impacted provider perceptions of the tool's ability to help them accomplish tasks quickly.

Additionally, as discussed above, it is possible that the evaluation study targeted the wrong decision-making process, or at least did not consider the full complexity of how providers might use the predictive model information to aid in decision-making. As the main focus was to examine the value of explanations, model information was only presented for single predictions and providers made decisions about individual patient cases. However, as noted by nurses in the qualitative inquiry study, risk trend information was considered useful in assessing changes in risk and was perceived as having higher clinical utility than single risk predictions. Moreover, physicians mentioned that the system would be useful in prioritizing patients, which suggests that an overview of risk predictions for a group of patients may be helpful in making prioritization decisions about patient groups. Because the system did not include risk trends or overviews of risk predictions for groups of patients, a vital piece of how the predictive model system might be used to make decisions in the clinical setting may have been missed. Finally, it's possible that the

evaluation metrics used in the evaluation study did not capture how explanations of predictive models might impact decision-making. While accuracy, confidence, and efficiency are obvious metrics to assess improved decision-making, it's possible that explanations of predictive models improve decision-making in subtler ways. Examples of alternative metrics of improved decision-making could include: 1) provider shared decision-making performance (e.g., explanations could facilitate conversations with patients that lead to better shared decision-making); 2) amount of information incorporated into a decision (e.g., explanations could prompt providers to view more patient information prior to making a decision, which could be viewed as beneficial as they would rely less on heuristic decision-making); and 3) effort required to make decisions (e.g., explanations may reduce cognitive effort required by providers to make decisions). Some of these metrics may be challenging to measure (e.g., shared-decision making performance[138]), but are worth considering for inclusion in future studies evaluating the potential impact of predictive model systems.

Despite the aforementioned limitations, provider perceptions of the predictive model with explanations were generally positive, indicating that they would accept and use the system in practice. These positive perceptions warrant further exploration of the pediatric ICU in-hospital mortality risk model with explanations to address some of the limitations and unanswered questions. Specifically, I would propose designing a system that presents the predictive model with the user-centered explanation as a tool integrated into the existing EHR. The system would incorporate user feedback on the displays of the supporting information, specifically mimicking the familiar EHR displays for laboratory test and vital sign data and incorporating the other suggestions for improvement provided by users. It would also include risk trend information for patients and provide overviews of risk predictions for groups or lists of patients. Studies assessing

the usability of the system could then be performed to gain a better understanding of how the system information fits into provider workflow and decision-making processes, specifically focusing on the utility of the user-centered explanation design. The combination of a "think-aloud" protocol analysis with "near-live" clinical simulations proposed by Li et al.[139] would likely be a good approach for these studies. The "think-aloud" protocol analysis would allow improvement of the usability of the system and give insight into how it might be used in clinical decision-making processes. The "near-live" clinical simulations would provide further information as to how the system could best accommodate clinician workflow and be used in the clinical setting, further elucidating the potential impact of the system on provider decision-making processes and patient outcomes. The results of this study could then be used to inform the design of evaluation studies of system impact.

Some other interesting directions for future work are suggested by the two provider viewpoints regarding the benefit of the predictive model with explanations. First, providers who viewed the predictive model as a confirmatory tool exhibited a level of distrust in the system, particularly when the model information contradicted their own knowledge. It is unclear from the study whether this distrust is why some providers view predictive models as confirmatory tools rather than as tools that could guide decision-making. This raises an interesting question as to how provider trust in a predictive modeling system may impact use of the system, and thus affect whether the system has an impact on outcomes. Future studies on how provider perceptions of trust in predictive modeling systems affects use of the system would be of interest. Second, it was interesting to note that some providers viewed the predictive model with explanations as a tool that could improve provider efficiency in prioritizing patients and reviewing patient information. Recent work has demonstrated improvements in patient information review efficiency when using

past provider viewing patterns to predict and highlight relevant information in the EHR.[140] In light of this work and viewpoint, an interesting future study might be to explore how instance-level explanations of models predicting clinical deterioration in patients could be used to effectively guide clinician review of patient information in the EHR by highlighting information of concern.

Another direction for future work arises from the finding that the SHAP explanations enabled providers to suggest ways to improve the clinical credibility and utility of the pediatric ICU in-hospital mortality risk model. This suggests that instance-level explanations could be useful communication tools for model developers to incorporate provider feedback and knowledge into models based on ML approaches. This could possibly involve interactive ML approaches in which healthcare providers use instance-level explanations to provide feedback on individual predictions to improve a model. Incorporating healthcare provider feedback and knowledge into models has been shown to improve acceptance of models in practice.[23,24] Feedback could be provided in a laboratory setting during model development or in the clinical setting as part of ongoing improvement of a deployed model. These settings would likely require different explanation needs and designs. While it was beyond the scope of this study, future studies could involve applying the theoretical framework to inform the design of explanations that facilitate provider involvement in the development and ongoing improvement of predictive models.

A final direction for future work would involve expanding the scope of the proposed theoretical framework. This could include adding components that extend the framework to account for how the use of specific data types or models might influence explanation design and interpretation. For example, to account for the influence of a specific model (assuming a model-agnostic explanation approach is not taken), a component could be added that demonstrates how knowledge of the specific model type (e.g., logistic regression, random forest) would influence

why the user might want an explanation as well as the space of possible explanation approaches and design options. Additionally, components could be included to provide more specific design suggestions based on the category of explanation approach (e.g., design options for model-agnostic, instance-level explanations of feature influence).

7.2 Conclusions

There is an increasing interest in high-performing predictive models capable of explaining the reasoning behind a prediction in a way that is both comprehensible and useful to healthcare providers. This dissertation aimed to address this need by proposing a new theoretical framework for user-centered explanation design of ML models in healthcare. The proposed framework was utilized in conjunction with healthcare provider feedback to inform the design of a user-centered explanation for predictions from a pediatric ICU in-hospital mortality risk model. While the user-centered explanation design improved provider perceptions of utilizing the predictive model in practice, the predictive model with the user-centered explanation did not demonstrate a significant improvement in provider accuracy, confidence, or efficiency in making decisions. Nonetheless, the work demonstrated that model-agnostic, instance-level, explanation approaches based on feature influence methods are a viable approach to explaining model predictions to healthcare providers. These explanations can be utilized for any model and can help overcome the model interpretability barrier to utilizing high performance ML models in practical applications in medicine. This work also identified several possible areas in which the proposed theoretical framework could be useful in designing explanations.

Overall, the work in this dissertation provides meaningful insights into the role of model interpretability and explanation in healthcare and contributes to knowledge on how to effectively communicate ML model information to healthcare providers. It is my hope that insights from this work can facilitate conversations with healthcare providers about the development, deployment, and continuous improvement of ML-based tools that can promote positive changes in clinical practice.

Appendix A Descriptions and Comparisons of SHAP and LIME Algorithms

This appendix provides a description of the LIME and SHAP algorithms and presents the experiments conducted to select between the two algorithms. Both algorithms generate explanations for a classifier or regressor in the form of feature-importance rankings and were developed to handle a variety of input data types, including image, text, and tabular data. Sections A.1 and A.2 provide an overview of how each algorithm works for a binary classification problem. I focused specifically on tabular data input as this is the most common data format used for risk prediction models in healthcare. Section A.3 presents the experiments conducted to justify the selection of the SHAP algorithm for use in this work.

Appendix A.1 Local Interpretable Model-agnostic Explanations (LIME)

The goal of the LIME algorithm[123] is to "identify an interpretable model over the interpretable representation that is locally faithful to the classifier".[64] In simple terms, to generate an explanation for a single instance, LIME uses a human-readable representation of classifier inputs (e.g., words instead of word vectors) to learn an interpretable model (e.g., sparse linear regression, short decision tree) that fits the local decision boundary near the instance of interest. Formally, for an instance x, family of interpretable models G, original predictive model f, and proximity measure Π_x, an explanation $\xi(x)$ can be produced by solving:

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \mathcal{L}(f, g, \Pi_x) + \Omega(g) \quad (1)$$

where $\mathcal{L}(f, g, \Pi_x)$ provides a measure of how unfaithful the interpretable model g is in approximating the original prediction model f in the locality defined by Π_x, and $\Omega(g)$ is a measure of the complexity of interpretable model g (e.g., tree depth for decision trees). To approximate $\mathcal{L}(f, g, \Pi_x)$ in a model-agnostic manner, LIME gains an understanding of the local behavior of the original predictive model f by generating perturbed samples, obtaining their predictions from f, and weighting them by their distance from the instance of interest (Π_x). Equation 1 can then be optimized to get an explanation by using the new weighted samples to fit an interpretable model that is constrained by the complexity parameter $\Omega(g)$. In practice, $\Omega(g)$ is a user-specified parameter indicating how many features to include in an explanation. Figure A1 provides a graphic depiction of the LIME approach to generating an explanation for a single instance. Currently, the implementation of LIME only supports explanations in the form of regressions. The exact approach to generating explanations varies by input data type, but an overview of the LIME implementation approach based on tabular data input is provided. Specific details on the implementation approaches for text and image data can be found in the LIME code and documentation.[141] The following implementation description is based on the code and documentation for LIME version 0.1.1.31.

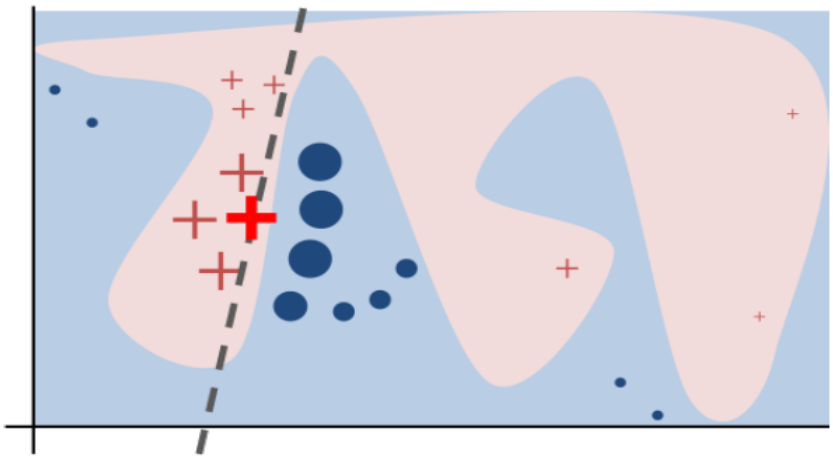


Figure A1. Graphic overview of the LIME approach to generating explanations. To generate an explanation for an instance of interest (indicated by the bolded red cross), LIME performs the following: 1) generates perturbed samples (all non-bolded points on the plot), 2) obtains their prediction from the classifier (circle or cross), 3) weights them according to distance from the instance of interest (represented here by point size), and 4) uses weighted samples to fit an interpretable model (indicated by the dashed line) that is locally faithful to the original predictive model decision boundary (indicated by the red/blue background). (Image taken directly from Ribeiro et al.[123])

As indicated in Figure A1, the first step in the LIME explanation process is to generate perturbed data samples. LIME requires training data to perform this step, which is usually the same dataset used to train the predictive model. Numerical features are perturbed by randomly sampling from the standard normal distribution and performing inverse mean centering and scaling using the feature means and standard deviations computed from the training data. Categorical features are perturbed by randomly sampling feature values according to their frequency in the training data, and then creating a binary feature to indicate whether the perturbed value matches the value for the instance being explained (i.e., 1 when the value matches, 0 otherwise). Users can specify the number of perturbed samples to generate for each explanation, but the default for tabular data is 5,000 samples. LIME then uses the original predictive model to obtain the class prediction

probabilities for each of the perturbed samples. Each perturbed sample z is weighted according to Π_x, which is defined as an exponential kernel:

$$\Pi_x(z) = \exp(-D(x,z)^2/\sigma^2) \quad (2)$$

where D and σ are a user-defined distance metric and kernel width, respectively. If not specified by the user, LIME will use Euclidean distance and a kernel width equal to 75% of the square root of the number of training data features. The weighted perturbed samples are then used to provide an approximation to Equation 1 by first selecting a specified number of features and then learning feature weights via regression. The user has control over the number of features selected, the approach to feature selection, and the type of regression. By default, LIME generates explanations using 10 features, selects features that have the highest product of absolute weight and original data point when learning a linear ridge regression with all features, and uses linear ridge regression with a regularization strength of α=1 to learn feature weights for the explanation. Users also have the option of discretizing numeric features in the explanation and are provided several discretization options. By default, LIME discretizes numeric features into quartiles for explanations.

As noted above, the implementation of LIME provides control over a variety of algorithm parameters. While this flexibility can be beneficial, the explanations produced can be heavily affected by the choice of these parameters. Defaults are provided for all parameters, but the LIME authors provide little guidance for parameter selection and do not provide justifications for the default settings.

Appendix A.2 SHapley Additive exPlanations (SHAP)

The SHAP algorithm[124,125] aims to unify several local explanation methods into a single approach for interpreting model predictions. It introduces the perspective of "viewing *any* explanation of a model's prediction as a model itself", and calls this model the *explanation model.*[125] Additive feature attribution methods are introduced as a class of explanation models that attribute an effect, ϕ, to each feature in a model and the sum of these effects approximates the original model prediction. The *explanation model* is thus defined as a linear function of binary variables:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z_i' \quad (3)$$

where z' is a binary vector of simplified features (e.g., binary vector indicating whether a specific feature was observed or not) of length M, and M represents the number of simplified features. This class of explanation model is used by several instance-level explanation methods, and therefore unifies these methods under a single approach. For specific details on each method and how it fits this class of explanation model, see Lundberg 2017.[125]

There exists a unique set of values of ϕ that ensures this class of explanation models meets three desirable properties: 1) local accuracy/fidelity (i.e., the sum of the attributed feature effects exactly equals the model prediction); 2) missingness (i.e., an absent input feature should have no attributed effect); and 3) consistency (i.e., if input feature always has greater impact in one model over another, then it should be attributed a higher effect for that model). This unique set of values are the Shapley values, a method from cooperative game theory that fairly distributes gains among all players of a collaborative game according to their marginal contributions towards the total gain.[142] For an explanation model, the "players" are the features, the "gains" are the effect

attributed to each feature, and the explanation model for a prediction, $f(x)$, can be formally defined as follows:

$$\phi_i(f,x) = \sum_{S \subseteq S_{all} \setminus \{i\}} \frac{|S|!(M-|S|-1)!}{M!} [f_x(S \cup \{i\}) - f_x(S)] \quad (4)$$

where ϕ_i corresponds to the Shapley value of the i-th feature, f is the original prediction model, x is the prediction instance to be explained, S is a subset of the set of all features except the i-th feature, $|S|$ is the number of features in the subset, M is the number of simplified input features, and $f_x(S) = f(x_S)$ where x_S is equal to the values in x for features in the set S but are considered missing otherwise.[75] As many prediction models do not support arbitrary patterns of missing input data, in practice $f_x(S)$ is estimated by computing its expected value on repeated evaluations of $f_x(S_{all})$ where missing values are filled in using randomly selected samples from a training dataset.[75] The ϕ_i of a feature can be interpreted as the change in the expected model prediction that occurs when a feature is observed versus unknown, averaged across all possible subsets and orderings of features.

To better clarify the theory and interpretation of the Shapley values produced by the SHAP algorithm, a simple example is provided. Imagine a model that predicts a person's risk of having the flu based on four features: 1) temperature, 2) presence/absence of a cough, 3) presence/absence of a runny nose, and 4) presence/absence of fatigue. Assume that the model predicts 0.10 probability of having influenza for the average person and the goal is to explain the prediction for a person who has a 0.75 probability in terms of how each of the four features impacts the model's prediction relative to the average person. Assume that this person has a temperature of 102.3°F, presence of a cough, no runny nose, and presence of fatigue. To find the impact of each feature on the prediction, their ϕ values must be computed as defined in Equation 4. A walk-through of each part of the calculation of ϕ for the presence of fatigue is described below and shown in Figure A2.

To calculate ϕ for the presence of fatigue, it is first necessary to estimate the marginal contribution of "fatigue = present" for each possible feature subset that can include that feature. For example, consider the last subset depicted in Figure A2 that includes the features "cough = present", "runny nose = absent", and "temperature=102.3°F". To estimate the marginal contribution of "fatigue = present" for this subset, model predictions are obtained when this feature value is observed and not observed (i.e., missing). When "fatigue = present" is observed, the model predicts a probability of 0.75. To get an estimate of what the model would predict if the value for fatigue was missing, a value for fatigue is randomly sampled from the training dataset. Assume a randomly sampled value of "fatigue = absent" and a model prediction of 0.60. Several repetitions of this sampling procedure can be performed, and the values can be averaged to obtain a better estimate of the model prediction when fatigue is missing. Assume the estimate after several repetitions was 0.60. Then, an estimated marginal contribution of "fatigue = present" for this subset would be 0.75 – 0.60 = 0.15. By taking a weighted average of the marginal contributions for each subset, an estimate of ϕ for "fatigue = present" is obtained (Figure A2).

Repeating the calculation shown in Figure A2 for each feature will yield the set of ϕs that comprise an explanation for the person of interest. A full explanation might read as follows: relative to the average risk prediction of 0.10, a temperature of 102.3°F increased this person's risk by 0.35, presence of a cough increased this person's risk by 0.20, absence of a runny nose decreased this person's risk by 0.05, and presence of fatigue increased this person's risk by 0.15. By summing all ϕ values with the average prediction, the person's prediction of 0.75 (i.e., 0.10 + 0.35 + 0.20 + -0.05 + 0.15 = 0.75) is obtained. Thus, the ϕ value of a feature provides an estimate of how much the feature changes the model prediction relative to the average prediction.

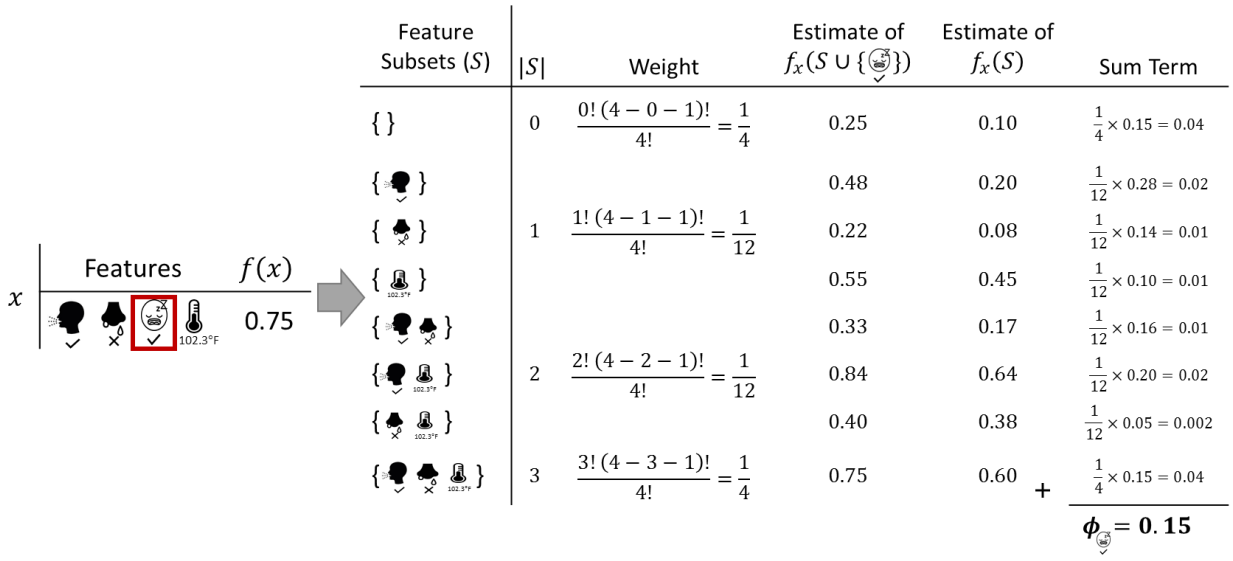


Figure A2. Calculation of the Shapley value, ϕ, for the presence of fatigue. All possible subset combinations are enumerated and weighted by the proportion of all possible feature permutations they represent. For each subset, the model prediction is estimated with and without the feature of interest, which are $f_x(S \cup \{fatigue = present\})$ and $f_x(S)$, respectively. These estimates are obtained by filling in any missing feature values with randomly sampled values from a training dataset and obtaining the model prediction, then averaging the predictions from repeated runs of this procedure. Subtracting estimates of $f_x(S \cup \{fatigue = present\})$ and $f_x(S)$ for a subset gives us an estimate of the marginal contribution of the presence of fatigue for that subset. The final Shapley value is obtained by taking a weighted sum of the marginal contribution estimates of each subset.

As can be clearly seen in the above example, the computation of the Shapley values for a set of features is non-trivial. The SHAP algorithm offers both model-agnostic and model-specific methods for efficiently approximating the Shapley values defined by Equation 4 to obtain explanations for any input data type. Although the authors point to previously defined model-agnostic methods for estimating Shapley values, they also include a new, more computationally efficient method called Kernel SHAP. The authors also provide computationally efficient, model-specific methods for estimating the Shapley values of linear models, deep learning models, and tree-based models. An overview of the Kernel SHAP explainer is provided below. Specific details

on the theory and implementation of the model-specific explainers can be found in the SHAP paper and code.[124,126] The following description for the Kernel SHAP explainer is based on the SHAP papers[124,125] as well as the code and documentation for SHAP version 0.24.0.[126]

Kernel SHAP proposes Shapley values as the solution to the linear model formulation of the LIME algorithm (see previous section), thus allowing for Shapley values to be approximated using a weighted linear regression. This permits a joint estimation of all Shapley values, which reduces the samples needed to provide accurate estimates of the Shapley values. To estimate the average model prediction and simulate missing features as in the example, Kernel SHAP requires a user-provided background dataset. This can be the entire training dataset used to learn the original predictive model; however, for larger datasets the algorithm becomes very computationally expensive. Therefore, it is recommended that for larger training datasets, users provide a dataset of reference values that adequately summarize the training data, such as point estimates for each feature (e.g., median or mean) or weighted samples produced by k-means or k-medians clustering. Kernel SHAP computes the average model prediction as the expected value of the model prediction on the background dataset. To efficiently estimate Shapley values, Kernel SHAP first begins by determining which feature values in the instance to be explained vary (i.e., have a different value) from the values in the provided background dataset. If a feature does not vary compared to the background dataset, it is assumed to have no effect on the model prediction and is assigned a Shapley value of 0. This helps reduce the number of computations required by the algorithm.

To estimate the Shapley values of the remaining features, Kernel SHAP first generates a weighted dataset of samples from all possible feature subsets, where features in the subset are equal to the value of the instance to be explained and features not in the subset are equal to the

background dataset values. Depending on whether the background dataset contains a single reference value or a set of reference values, a "sample" in the weighted dataset may consist of a single row or a set of rows, respectively. To further reduce computation time, users can specify the number of samples (i.e., model evaluations) that Kernel SHAP is permitted to use, with higher sample sizes leading to more stable estimates. By default, Kernel SHAP uses $2^{(\#\text{ of varying features})} + 2^{11}$ samples and caps the maximum number of samples allowed at 2^{30}. If given enough samples, Kernel SHAP will fully enumerate all possible subset sizes; otherwise, the algorithm first enumerates as many high-weighted subset sizes as possible (e.g., $|S| = 0$ and $|S| = 1$ in Figure A2), then uses any leftover samples to randomly sample subsets from the remaining subset sizes. If more samples are allowed than are needed to fully enumerate each subset, unused samples are discarded to improve computational efficiency. For each of the samples in the weighted dataset, the algorithm estimates the change in the model prediction from the average model prediction. For a background dataset using a set of reference values, this estimate is the expected value of the model prediction over all rows in the sample minus the average model prediction. By default, if less than 20% of all possible subsets have been enumerated in the weighted dataset, Kernel SHAP performs feature selection using a Lasso model with least angle regression using the Akaike information criterion for model selection. The user has optional control over whether to run feature selection as well as the L1 regularization parameter used in the Lasso model. Finally, Kernel SHAP uses the weighted dataset of samples and their respective estimated changes in the model prediction to solve a least squares regression to obtain the Shapley values for the remaining features.

As with the LIME algorithm, the implementation of the SHAP algorithm allows users control over parameters that could impact the explanations generated. Thus, parameters require careful selection by the user.

Appendix A.3 Algorithm Comparison Experiments

To select between the SHAP and LIME algorithms, I performed experiments comparing the algorithms on two properties of explainers previously identified as desirable in the literature: 1) fidelity (i.e., the explanation model should accurately reflect the underlying predictive model's behavior) and 2) computational efficiency.[12,20,21,65] I believe that these two properties will be essential for any explanation approach used in healthcare. For an model-agnostic, instance-level explanation approach based on feature influence, I proposed the following metrics for quantitatively measuring these properties:

Fidelity: There should exist some function of the set of generated feature influence values, $g(\phi)$, that approximates the original model prediction, $f(x)$. A high-fidelity explanation approach is one that generates explanations such that $g(\phi) \cong f(x)$.

Computational Efficiency: The time required to generate a single explanation.

Several preliminary experiments comparing the LIME and SHAP algorithms on their fidelity and computational efficiency were conducted. To conduct experiments, two datasets curated in previous research projects were used: 1) a dataset to predict 30-day all-cause pediatric hospital readmission risk and 2) a dataset to predict 1-year postpartum infant mortality risk. To enable compatibility with the explanation algorithms, comparable Python versions of the models previously learned on each of these datasets were generated. A short Python module was developed

to facilitate the use of both explanation algorithms and conduct experiments. Datasets and models are described in subsection A.3.1, the experiments are described in A.3.2, the results are presented in subsection A.3.3, and subsection A.3.4 presents a discussion of all results and uses them to justify the selection of an explanation algorithm to be used in the work.

Appendix A.3.1 Datasets and Models

30-day all-cause pediatric hospital readmission risk: This dataset constituted all clinical and administrative data for all inpatient visits to CHP from January 1, 2007 to December 31, 2013. Patients that died during admission, were over 21 years of age, or did not have a recorded age were excluded from the dataset. A readmission was defined as any inpatient visit followed by a second inpatient admission within 30 days of discharge from the initial visit. Multiple readmissions within 30 days for a single patient were treated as separate cases. The final dataset was comprised of 91,045 visits (13,548 readmission cases; 77,497 non-readmission controls).

For brevity, I have left out the specific details of the data cleaning, standardization, and feature engineering processes. It is important to note that all numeric features were discretized using the minimum-description-length criterion discretization method[101] and missing data were treated as separate categories. In the original model learning process, features were selected using a two-stage predictor-selection process which included an IG filtration step followed by a wrapper-based search. A series of different Naïve Bayes models using various combinations of medical data sources (e.g., medications+labs, demographics only, etc.) were trained using WEKA.[103] The best performing model was trained using all data sources, included 32 features, and achieved an AUROC of 0.806 on an independent test dataset. To generate a comparable Python version of this model, a 70/30 stratified split of the 91,045 visits in the dataset was used to generate training and

testing data. Using the 32 features from the best performing model and one-hot encoding procedures, a Naïve Bayes classifier was learned on the training data. The learned model achieved performance comparable to the original best performing model, with an AUROC of 0.806 for the test dataset.

1-year postpartum infant mortality risk: This dataset was obtained from the Magee Obstetric Medical and Infant (MOMI) Database and comprised demographic and medical information for all deliveries at Magee-Women's Hospital from January 1, 2002 to December 31, 2014. Infant death cases were identified by linking the MOMI dataset with data from the Department of Health Services (DHS). Stillbirths and fetal deaths were excluded. The final dataset encompassed 75,842 records (494 infant death cases; 75,348 alive controls).

Again, for brevity, I have left out the specific details of the data cleaning, standardization, and feature engineering processes. It should be noted that missing values were imputed by simple random sampling from known data and continuous variables were discretized using the entropy minimization heuristic method. The final dataset contained 102 features for analysis. A variety of models were learned using R, but the highest performing model was a ridge logistic regression trained on 29 features selected using a sequential IG filter, which achieved an average AUROC of 0.933 with 10-fold cross-validation on the training dataset (i.e., all 75,842 visits). To generate a comparable Python version of this model, all 75,842 visits in the dataset and one-hot encoding procedures were used to learn a ridge logistic regression using the 29 features from the best performing model. The learned model exhibited comparable performance to the original best performing model, achieving an average AUROC of 0.925 with 10-fold cross-validation on the entire dataset.

Appendix A.3.2 Experiments

500 patients (250 cases, 250 controls) were randomly sampled from each dataset to use in experiments. As noted in sections A.1 and A.3, the parameter settings can affect the explanations produced by both the LIME and SHAP algorithms; therefore, for each patient, LIME and SHAP explanations were generated under varying parameter settings. For the LIME algorithm, varied parameter settings included: 1) the number of perturbed samples used to learn the linear regression model and 2) the number of features selected for the explanation (i.e., the two parameter settings most likely to influence the explanations generated). For the SHAP algorithm, a background dataset consisting of the median value for each feature was used and varied parameter settings included the number of samples used to estimate the Shapley values of the features. Default values for all other user-controllable parameters were used. Explanations were generated for the prediction of the target class of interest (i.e., "Readmitted" for readmission model and "Death" for infant mortality model). The two properties identified in the introduction were as assessed as follows:

Fidelity: Following the measure defined in the introduction, fidelity error was estimated for each explanation as $g(\phi) - f(x)$, where $g(\phi)$ is a function of the set of generated feature influence values that approximates the original model prediction, $f(x)$. For the LIME algorithm, $g(\phi)$ takes the form of a local linear regression. For the SHAP algorithm, $g(\phi)$ is simply the sum of all the Shapley values learned for each feature and a base value (i.e., the average model prediction or the expected model prediction when no features are known). Theoretically, the SHAP algorithm guarantees fidelity (i.e., guarantees that $g(\phi) = f(x)$), but as approximation methods are used to estimate the Shapely values it is beneficial to check that this guarantee holds true for

the implementation of the algorithm. For each dataset, the median absolute error (MAE) in fidelity was calculated across all 500 patients for each algorithm and varied parameter setting.

Computational Efficiency: To estimate computational efficiency of the algorithms, the time to generate each explanation was measured. Mean and total computation times for each algorithm were calculated across the 500 patients from each dataset under varying numbers of samples used to generate the explanation. For the LIME algorithm, the number of explanation features had minimal impact on computation time for a single explanation and so all timing experiments were performed using 6 features in the explanation. This value was based on the findings from the fidelity experiments (i.e., LIME's fidelity error on the two datasets appears to be optimal somewhere between 5 and 10 features).

Appendix A.3.3 Results

Figure A3 shows the MAE in fidelity on each dataset for the LIME algorithm under varying parameter settings. The MAE in fidelity for the SHAP algorithm was always 0 for both datasets.

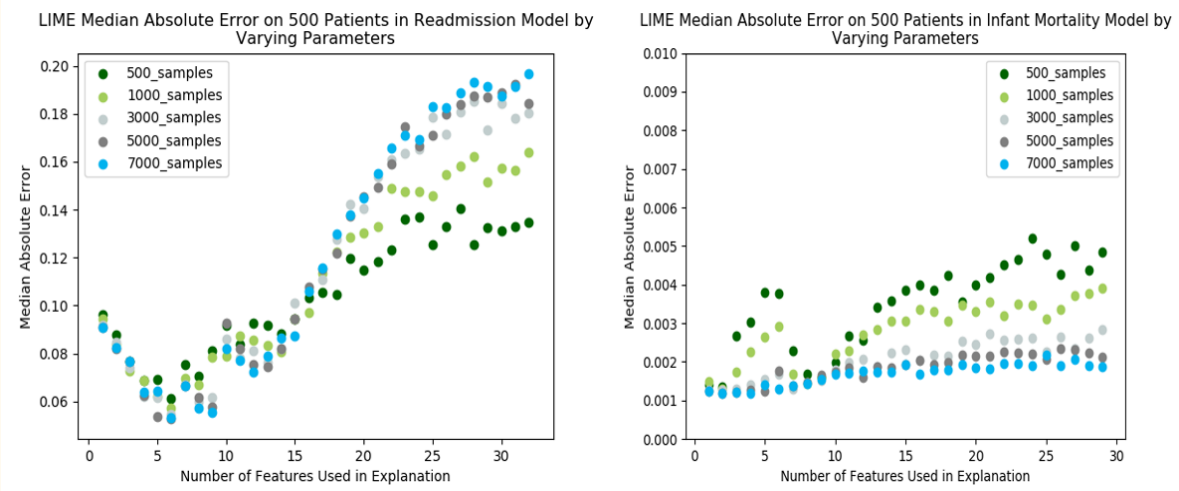


Figure A3. LIME median absolute error (MAE) in fidelity. The MAE in fidelity for the LIME algorithm under varying parameter settings is shown for the readmission dataset (left plot) and the infant mortality dataset (right plot).

The mean and total computation times for the LIME and SHAP algorithms under varying parameter settings are shown in Table A1 and Table A2, respectively.

Table A1. Mean time to compute a single explanation for LIME and SHAP algorithms

		Number of samples used to generate explanation				
		500	**1000**	**3000**	**5000**	**7000**
Readmission Dataset	**SHAP mean time (s)**	0.04	0.05	0.09	0.11	0.14
	LIME mean time (s)	0.39	0.54	1.26	1.90	2.52
Infant Mortality Dataset	**SHAP mean time (s)**	0.05	0.05	0.04	0.01	0.01
	LIME mean time (s)	1.09	1.53	3.90	1.83	2.38

Table A2. Total time to compute 500 explanations for LIME and SHAP algorithms

		Number of samples used to generate explanation				
		500	**1000**	**3000**	**5000**	**7000**
Readmission Dataset	**SHAP total time (min)**	0.31	0.46	0.75	0.92	1.16
	LIME total time (min)	3.26	4.50	10.53	15.8	21.0
Infant Mortality Dataset	**SHAP total time (min)**	0.45	0.44	0.33	0.07	0.06
	LIME total time (min)	9.12	12.76	32.54	15.25	19.8

Appendix A.3.4 Discussion and Algorithm Selection

The fidelity error of the SHAP algorithm under varying parameters was always 0 for both datasets, which indicates that the algorithm implementation adheres to its theoretical guarantee of local fidelity. On the other hand, Figure A3 demonstrates that the LIME algorithm error in fidelity varies across parameters and datasets. This indicates that use of the LIME algorithm to generate explanations would require careful selection of algorithm parameters for each dataset to reduce errors in fidelity. Additionally, as no parameter setting is likely to be ideal for all instances in a dataset, it would be necessary to show users an estimate of the error in the explanation generation process. Thus, the SHAP algorithm seems to be a better choice to ensure explanation fidelity.

Table A1 and Table A2 show that the SHAP algorithm appears to be faster than the LIME algorithm and its computation time is less affected by the number of samples used to generate the

explanation. It should be noted that the SHAP algorithm computation time is highly dependent on the background dataset provided to the algorithm. Although not explored in these preliminary experiments, larger background datasets may lead to significant decreases in the algorithm's computational efficiency. However, unlike the LIME algorithm where each explanation must be computed individually, the SHAP algorithm also includes functions to efficiently compute explanations in large batches. These functions were not explored in the preliminary experiments but are worth noting for future studies. Based on these preliminary timing experiments, the SHAP algorithm appears to offer better computational efficiency than the LIME algorithm.

As the SHAP algorithm guarantees explanation fidelity and requires less computation time than the LIME algorithm, the SHAP algorithm was selected for use in the proposed work. It is important to note that only preliminary experiments were conducted; however, the conducted experiments provided sufficient evidence to support my selection of the SHAP algorithm. More rigorous experiments comparing the LIME and SHAP algorithms are left for future work.

Appendix B Qualitative Inquiry Questionnaires and Question Guide

Background Questionnaire

1. Please select the choice that most accurately describes your position at Children's Hospital of Pittsburgh (CHP).

- ○ Attending
- ○ Care manager
- ○ Fellow
- ○ Intern
- ○ Nurse
- ○ Nurse practitioner
- ○ Physician assistant
- ○ Resident
- ○ Other (please specify): _____

2. Approximately how long have you been practicing clinically?

3. Please **check all** statements that apply regarding your level of familiarity with risk prediction models.

- ☐ I know what a risk prediction model is.
- ☐ I have used a risk prediction model in practice.
- ☐ I have been involved in the development of a risk prediction model.

4. Please indicate your level of familiarity with the field of machine learning.

- ○ None -- I have never heard of this term before.
- ○ Basic awareness -- I have heard the term, but don't know much about it.
- ○ Know a little -- I am familiar with the main concepts of machine learning.
- ○ Know a fair amount -- I have a practical understanding of machine learning concepts
- ○ Know it well -- I have a theoretical understanding of machine learning concepts

5. Please indicate your level of agreement with the following statements:

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
Predictive analytics will play an important role in the future of medicine.	○	○	○	○	○
There are too many complexities and subtleties in medical practice for computers to be helpful in predicting clinical outcomes.	○	○	○	○	○
I have reservations about using predictive analytics solutions in clinical practice.	○	○	○	○	○

Focus Group Question Guide

Model Discussion
Question Guide • *How would you feel about deploying these kind of predictive models into clinical practice?* • *What practical applications do you think these kinds of models could have in clinical practice?* • *Would you feel confident in the predictions provided by these kinds of models?*   • *What additional information about the model would you require in order to have confidence in its predictions?* • *Do you think you would use predictions from models like this? Why?* • *Apart from predicting other outcomes, how could these kinds of models be made more useful?* • *You may have noticed that not much information about the model or the underlying algorithm was provided. How might this information influence your perceptions of a model? What assumptions, if any, did you make about the model or underlying algorithm?*
Mock-up Review

Individual Mock-up Question Guide
• *How would you summarize why the model made this prediction?*
• *Why might you be inclined to believe or disbelieve a prediction presented in this fashion?*
  • *Are any predictors surprising or non-sensical?*
  • *What information led you to belief/disbelief of the prediction?*
• *What information is missing that might help you interpret this prediction more effectively or efficiently?*
  • *Model performance? Confidence intervals for contribution values?*
  • *Different grouping or order of predictors? Different number?*
• *What information provided might you find useful in performing your job?*
Mock-up Set 1 Comparison Question Guide
• *What do you think of displaying risks as probabilities versus odds? Which do you prefer? Why?*
• *What do you think of displaying individual predictors versus groups of predictors? Which do you prefer? Why?*
• *What do you think of the tornado plot versus the force plot? Which do you prefer? Why?*
• *What would you change about any of these displays?*
  • *What changes would make a display easier to understand?*
  • *What information or design elements do you think are missing?*
  • *What information or design elements are not useful?*
Mock-up Set 2 Comparison Question Guide
• *How does grouping predictors into plots change your opinion of displaying individual predictors versus groups of predictors?*
• *What do you think of grouping predictors into multiple explanation plots?*
  • *What alternative ways to group predictors can you think of?*
  • *What is your preferred grouping, or would you prefer no grouping?*
• *What would you change about any of these displays?*
  • *What changes would make a display easier to understand?*
  • *What information or design elements do you think are missing?*
  • *What information or design elements are not useful?*

Design Characteristic Preferences and Ranking

The mock-ups we just reviewed varied 5 different aspects, or design characteristics, of a prediction explanation display. The design characteristics covered today included: 1) unit of explanation, 2) explanation size, 3) organization of explanation units, 4) explanation display format, and 5) risk representation. You were asked to review several different variations, or options, for each characteristic. In the sections below, you will be asked to indicate your option preferences for each design characteristic and rank the importance of each characteristic.

Section A.

For each of the design characteristics below, use the blank boxes to rank each option in order of preference (1—most preferred). Please assign a rank to each option and avoid assigning the same rank to different options for the same design characteristic.

Unit of explanation

☐ Individual predictors ☐ Predictor summaries (e.g., "Peds Coma Score results")

Risk representation

☐ Probability ☐ Odds

Explanation display format

☐ Force plot ☐ Tornado plot

Organization of explanation units

☐ No grouping ☐ Grouped by assessment (e.g., demographics, labs, vitals)

☐ Grouped by influence on prediction (e.g., factors increasing risk, factors decreasing risk)

Explanation size

☐ Modifiable

☐ Static

Section B.

Rank each design characteristic in order of perceived importance (i.e., rank 1 would indicate the design characteristic you felt was most helpful in understanding the model prediction). Please assign a unique rank (1-5) to each design characteristic.

☐ Unit of explanation ☐ Explanation display format

☐ Explanation size ☐ Risk representation

☐ Organization of explanation units

Appendix C Qualitative Inquiry Codebook

Name	Description
1. Context of use--when & where	The environment in which the explanation will be used, which is often related to the stage of system development. Environment will dictate the available user time and cognitive capacity, the available technical resources, and the user's perception of the system, which all may influence explanation design. This main category code is meant for organizational purposes only and should not be applied.
1.1 Environment	Aspects of the environment that will affect how an explanation needs to be designed in order to support use within that environment. This parent code should only be applied when a participant comment falls within this parent category, but none of its children codes can be applied appropriately.
1.1.1 Cognitive and time resources	Participant cognitive capacity and/or time availability to use the system in a specific environment. This includes comments about cognitive effort to process information in a given time frame (e.g., ease and speed of information processing, time restrictions, willingness to spend mental effort or time) and workflow or other environmental influences that imply a possible impact on cognitive capacity or time availability (e.g., task order, when/how/where to capture attention) Example: -Speed or ease with which knowledge can be obtained from system (e.g., faster synthesis of relevant information, familiarity with the way information is presented)
1.1.2 Social and organizational influences	Any aspect of the social or organizational environment in which the system is being used that may impact system development, design, or application. This can include things related to participant workflow, organizational infrastructure (e.g., staffing procedures/challenges, patient triage/bed assignment procedures/challenges, education/training programs, financial policies), and social pressure/expectations. Examples: -Workflow, such as rounding practices, patient/colleague interactions, EHR interactions, etc. -Staffing/triaging procedures, such as bed availability, staff availability, etc.
1.1.3 Technical resources	Technical resources available (e.g., compatibility with existing systems, processing/memory constraints) when using the system in a specific environment. This can include limitations of pre-existing systems, difficulties with real-time data processing, and challenges in implementation and/or maintenance of the system.
1.2 System stage	Design/information needs in a specific system stage (e.g., development, implementation, deployment). This code should only be applied when a different system stage may require a change in information/design needs. Example: -a participant mentions specific information which would assist in validating the predictive model (this may be an indirect reference to information/design needs in the development stage, which may differ from needs in the deployment stage)

2. Context of use--who	User's cognition (e.g., knowledge, experience, capabilities, etc.) and the user's relationship to the system at the time the explanation is being provided. A user may have several different relationships with the system over time, and thus their explanation needs may change with varying roles. This main category code is meant for organizational purposes only and should not be applied.
2.1 Cognition & experiences	The knowledge, experience, capabilities, etc. of the user of a system. Three main categories of user cognition to consider include AI experts, domain experts, and lay persons. Of particular interest is any aspect of the user's background knowledge or prior experiences that may bias their opinion of or attitude toward a new system. This parent code should only be applied when a participant comment falls within this parent category but none of its children codes can be applied appropriately.
2.1.1 Background knowledge	Participant's prior level of knowledge of ML/AI/predictive modelling concepts. Includes remarks/questions that suggest knowledge (or lack thereof) of predictive models (e.g., objective reference to known model, mentions of ML algorithms, model limitations/validity) or the model development process (e.g., cohort definition, data cleaning, feature engineering, training, evaluation, bias/overfitting, best practices). Not applicable to remarks/questions on presentation content (e.g., AUC, inputs)
2.1.2 Prior experiences	Participant's prior experience with using an ML/AI tool or another information system (e.g., EHR). This code is restricted for use when the participant expresses either a positive or negative opinion or attitude about the design, credibility, usability, or utility of the tool/system, and should not be used to code objective comparisons of tools/systems (e.g., comparing performance or data inputs, objective discussions on design and implementation).
2.2 Relationship to system	The user's relationship to the system at the time the explanation is being provided. Main roles to consider can be engineer, developer, owner, end-user, data subject, and stakeholder. It should be noted that a user may occupy more than one role simultaneously. This parent code should only be applied when a participant comment falls within this parent category but none of its children codes can be applied appropriately.
2.2.1 User perspective	How system design or system application might differ based on the user's current relationship with the system. This can include comments about design differences based on intended system application (e.g., developer vs end-user needs, different end-user information needs). This code should not be applied to design/application differences that would arise from variation in user cognition & experiences (e.g., background knowledge, thought processes)
3. Context of use--why	User needs and goals that drive the need for an explanation. Four general reasons why explanations of intelligent systems are required include verification, improvement, learning, and compliance. Needs/goals will vary according to the who/when/where elements of the context of use. This main category code is meant for organizational purposes only and should not be applied.
3.1 Compliance	Closely related to verification (3.4), this refers to any activities aimed at ensuring the system adheres to an established legal, moral, or other societal standard. This includes all comments on the system from an ethical, moral or legal/organizational policy standpoint.

3.2 Improvement	Closely tied with verification (3.4), this covers activities related to improving system performance and efficiency. May include incorporating domain knowledge to reduce biases in or improve generalization of model, comparing/selecting models, and improving system response times. Includes suggested changes in data collection, data inclusion/exclusion, data processing, and target outcomes/definitions. Not applicable to comments about or suggestions to improve model utility (e.g. possible applications of current model information, post-hoc analyses such as distribution of risk scores across units, or tracking risk scores and outcomes of specific patients). Examples: -suggestions to explore predictions of an outcome other than 24-hr pediatric ICU mortality (e.g., time to event predictions, morbidity, ICU transfer, mortality in a specific patient population, etc.) -suggestions to include additional data such as tests, staffing, comorbidities, bed location, etc. -suggestions to improve data processing such as removing outliers, dropping bad values, defining normal ranges, adjusting for age/condition, etc.
3.3 Learning	Remarks/questions indicating participant is seeking to gain knowledge or information from the system, including identifying new data patterns, generating/testing new hypotheses, and/or providing support for decision-making (e.g., provide supporting evidence for a decision, improve decision-making speed or accuracy, identifying actionable information such as courses of action or modifiable risk factors).
3.4 Verification	A possible reason for requiring an explanation of an intelligent system. Includes examining how decisions/suggestions are made by the system to ensure it is operating as expected, which may include activities such as detecting biases, finding/debugging errors, and ensuring that system reasoning aligns with domain knowledge. This parent code should only be applied when a participant comment falls within this parent category but none of its children codes can be applied appropriately.
3.4.1 Comparing against known model	Comparison of model to existing model/tool to validate some aspect of the system (e.g., credibility). Only applicable to remarks that compare objective metrics (e.g., performance, data collection & processing). Not applicable to participant opinions of existing models or preferences for system content & design (use "prior experiences" and "explanation design" category codes instead). Generally not applicable to comments related to system utility (use "Learning" category codes instead).
3.4.2 Comparing model information to domain knowledge	Comparison of system information against clinical knowledge to verify some aspect of the model (e.g., credibility). This can include comments or questions about possible data biases, information validity, etc. Not applicable to remarks where participants are suggesting improvements based on clinical knowledge. Generally not applicable to comments related to system utility (use "Learning" category codes instead).
3.4.3 Seeking information on model development processes	Remarks/questions seeking to validate any aspect of the model development and maintenance process (e.g., cohort definition, data sources, data collection/inclusion/exclusion, cleaning processes, feature engineering/selection, model learning process, evaluation, maintenance over time). Applicable only when participants make assumptions about or attempt to clarify/understand/question the model development/maintenance process and is not applicable to suggestions for improvement.

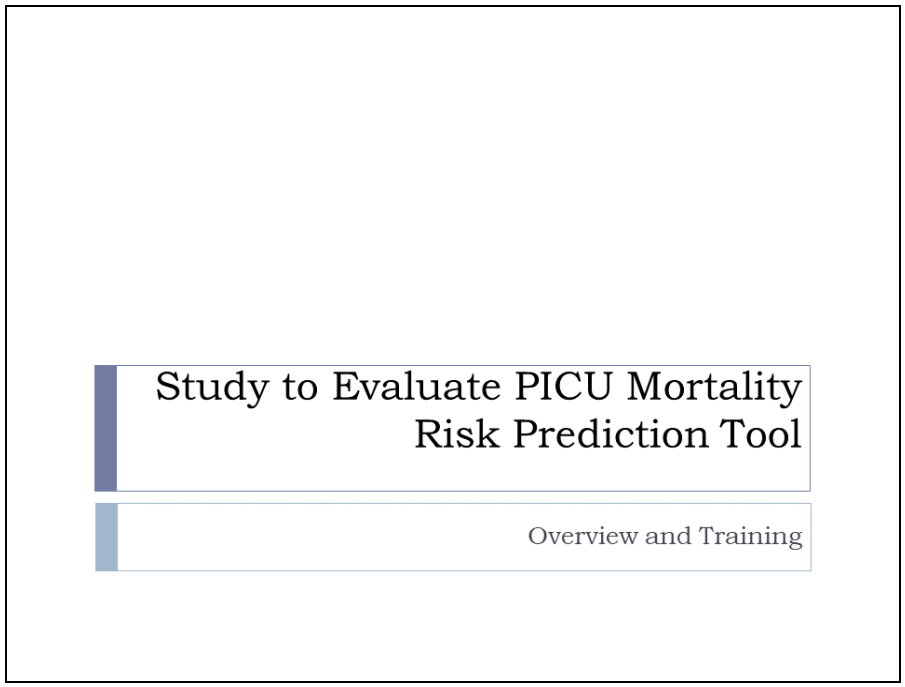
4. Explanation design--how	Can generally be determined by the who and why questions of context of use, and refers to the way in which the content of an explanation is presented to a user. The presentation of an explanation can generally be summarized using 3 main categories: dimensionality, explanation unit granularity and organization, and information representation. This main category code is meant for organizational purposes only and should not be applied.
4.1 Dimensionality	A main category to consider when designing an explanation. Refers to the processing size/levels of explanation information, which may include the overall size of an explanation or interactive exploration options. Should only be applied when a participant comment falls within this parent category but none of its children codes can be applied appropriately.
4.1.2 Size & interactivity preferences	Preferences for the size and/or interactivity options in the explanation design. Applicable to preferences on interactivity/size options in mock-ups (e.g., plot hover and drop-down select capabilities, link between explanation plot and predictor table, scrollable list of predictors) and any suggestions for interactivity/size options not shown in mock-ups (e.g., interactions between visualizations, amount of information content, interactive explanation exploration options).
4.2 Explanation unit & organization	Preferences regarding the granularity and organization of the explanation units. This includes preferences on the unit of explanation or predictor granularity (e.g., raw predictors, grouped/summarized predictors, increasing/decreasing or net contributions) and organization of the explanation units (e.g., order of display, location of increase/decrease contributions, grouping into different plots). Applicable to both remarks on mock-up options and suggested alternatives.
4.3 Information representation	A main category to consider when designing an explanation. This includes the vocabulary, data structures, and visualizations used to express information. This parent code should only be applied when a participant comment falls within this parent category but none of its children codes can be applied appropriately.
4.3.1 Data visualization preferences	Specific preferences for how data is displayed in the explanation design, which includes data structures (e.g., free-text, data tables, lists) and graphical representations (e.g., images, plots/charts, diagrams) used to display information. This includes participant preferences for mock-up options (e.g., tornado vs. force plot) and alternative suggestions. Applicable to participant suggestions for new or alternative displays. Generally not applicable to information content preferences or vocabulary/phrasing preferences, use "explanation design--what" and “vocabulary preferences” codes instead.
4.3.2 Vocabulary preferences	Specific preferences for the vocabulary used in the explanation design. Includes how test content is worded (e.g., phrasing used to describe predictors and contributions), expression of numerical information (e.g., risk in probability vs. odds, displaying probability as decimal or percentage), and domain-specific terms/abbreviations that should be used. Often applicable when participants express confusion/difficulties when trying to interpret text/numerical information.
5. Explanation design--what	Generally determined by the answers to the who and why of the context of use, and refers to the content that needs to be included in an explanation. Content of an explanation typically refers to the type of explanation being provided and any information supporting the interpretation of that explanation. This main category code is meant for organizational purposes only and should not be applied.

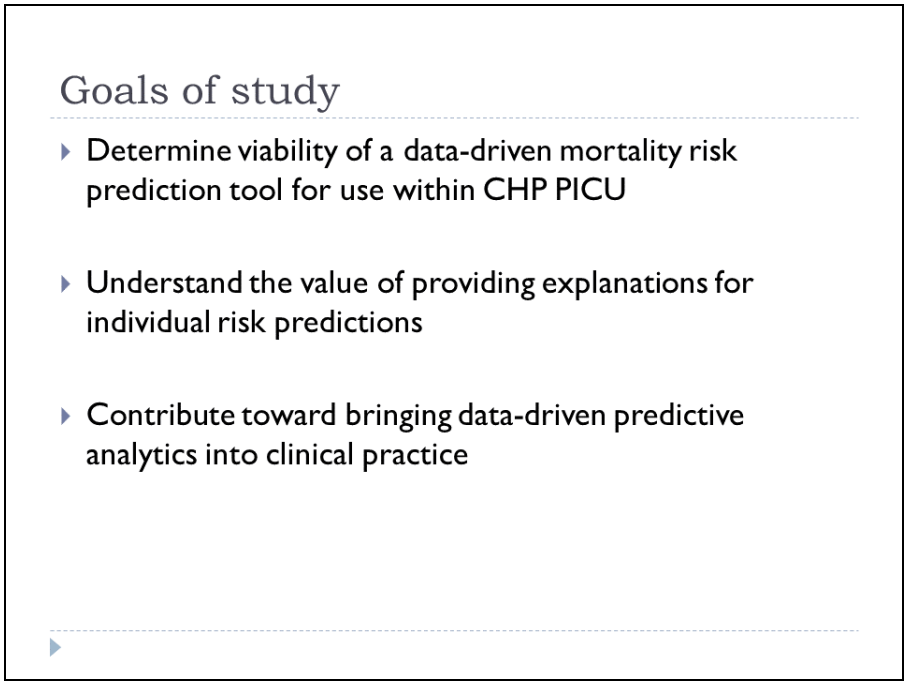
5.1 Supporting information	Any information that is not a part of the explanation but is required to help support the user's interpretation/understanding of the explanation. This may include things such as source data used in the model or explanation algorithm, supplemental data, and training materials. This parent code should only be applied when a participant comment falls within this parent category but none of its children codes can be applied appropriately.
5.1.1 Interpretation information	Needs for training information on how to interpret explanation information. This includes remarks/questions that indicate participant confusion and/or lack of understanding based on the system design (e.g., trouble interpreting predictors). Not applicable to momentary confusion (i.e., if participant voices question but quickly figures it out themselves). Not applicable to suggestions for data to include in interface to support explanation interpretation (use other "source & supplemental data" code instead). Not applicable to preferences/opinions on system design. Examples: -confusion on how to interpret predictor descriptions (e.g., making sense of discretized ranges or feature descriptions) -confusion on how to interpret predictor contributions and their relation to the baseline and model predictions Examples where "interpretation information" and "source & supplemental data" (5.1.2) both apply: -If it was more clear how to interpret xxx information, the xxx information would help me better understand the prediction and/or explanation -XXX information seems like it might be useful in understanding the prediction and/or explanation, but I find it confusing to interpret -If the system could include xxx information expressed in yyy manner, it would really help me interpret/understand the prediction/explanation
5.1.2 Source & supplemental data	Preferences/suggestions for including information about the prediction model (e.g., performance statistics, certainty measures, development processes), source data used by the model or explanation algorithm (e.g., raw data used to derive predictors, (un)discretised predictor values, contribution values), or any other supplemental data required to support interpretation of the prediction or explanation (e.g., interventions, care context). Not applicable to suggestions for improvements to the model. Examples: -direct/indirect comments on utility of information in explanation plot, predictor table, raw data plots, etc. (e.g., participant uses raw data plot or predictor table to investigate a predictor in explanation plot) -requests for information on model (e.g., confidence intervals, performance information, feature engineering/selection, etc.) -requests for information not used by model such as care interventions performed, staffing/triaging/bed assignment procedures that may have affected care, additional patient data needed to interpret prediction, etc. -comments on utility of diagnosis, demographic & utilization tables Examples where "interpretation information" (5.1.1) and "source & supplemental data" both apply: -If it was more clear how to interpret xxx information, the xxx information would help me better understand the prediction and/or explanation -XXX information seems like it might be useful in understanding the prediction and/or explanation, but I find it confusing to interpret -If the system could include xxx information expressed in yyy manner, it would really help me interpret/understand the prediction/explanation

5.2 Type of explanation	One part of explanation content refers to the type of explanation that is required, such as whether the explanation is one of processes or behavior and whether it is targeted at the local or global level. Type of explanation can generally be determined by the type of questions the user is asking or the reasoning processes the user is trying to use. This parent code should only be applied when a participant comment falls within this parent category but none of its children codes can be applied appropriately.
5.2.1. Intelligibility query	Specific intelligibility queries (i.e., “inputs”, “outputs”, “certainty”, “why not”/”how to”, “why”, “what if”, “when”) about the system. Includes comments indicating desire to know what data/predictors/inputs are used, what predictions/outputs can be produced, how (un)certain the model is in its predictions, why inputs produce certain outputs or how to get specific outputs, how changing inputs influences outputs, etc. Coding for intelligibility queries in the form of a question should generally not include answers to the question. Often applicable when "seeking information on model development process" code is used. Generally, not applicable to remarks regarding specific design elements. Examples: -questions on data/inputs being used by the model -comments/questions on predictions, including certainty of predictions, how/why predictions are produced, how changes in inputs might influence predictions.
5.2.2 Level & target preferences	Preferences for explanation level (local/global) and target (behavior/processes). Includes comments/questions directly/indirectly expressing an interest in knowing model internals (e.g., weights, mathematical relationships, handling correlated predictors), general trends learned (e.g., how/why the model makes predictions for patient population; general risk factors), and how/why the model makes predictions for individual patients (e.g., patient-specific risk factors).
6. Perceptions of the system	Perceptions of the overall system application. This includes perceptions on the barriers and facilitators to system adoption. For risk prediction models, adoption is closely tied to the utility, credibility, and usability of a model or system. This parent category code should only be applied when a participant comment falls within this parent category but none of its children codes can be applied appropriately.
6.1 Perceptions of system credibility	The credibility, or "believability", of the system. Includes comments on any aspect of the system that may influence the participant’s confidence in prediction accuracy (e.g., high performance may increase confidence, predictors that are outliers/bad data points may decrease it). Not applicable to remarks about the credibility of existing systems, use "prior experiences" code instead. Often applicable with "verification" category codes. Examples: -willingness to use/trial system based on performance (e.g., AUC) -scepticism about model predictions based on identified data errors/biases, missing info, etc. -comparing model performance/content with domain knowledge or to known model
6.2 Perceptions of system usability	The usability, or ease of use and learnability, of the system (i.e., can the intended goal be accomplished using the system or will users have difficulty?). Includes comments about aspects of the system that facilitate or impede use (e.g., design elements that make information processing easier/harder) and preferences between mock-ups (e.g., saying one mock-up was easier to use/understand than another). Not applicable to remarks about

	the usability of existing systems, use "prior experiences" code instead. Often applied with “explanation design” codes. Examples: -about mock-ups that are easier/harder to use than others -design elements that exacerbate/alleviate confusion, cognitive effort, time requirements -design elements that facilitate information synthesis or interpretation
6.3 Perceptions of system utility	The utility, or usefulness, of the system (i.e., is the intended use of the system useful to pursue? will users use it?). Includes suggestions for possible users of the system and comments on the value of system information (e.g., information provided is perceived as informative). Not applicable to remarks about the utility of existing systems, use "prior experiences" code instead. Often applicable with "learning" & "improvement" codes. Examples: -suggesting possible applications of the current model/system -(dis)interest in continued development of system -(dis)interest in information provided by system (e.g., “it’s not telling me anything new”, “this information could support xxx decision or help me determine xxx faster”)

Appendix D Evaluation Study Introductory Slides





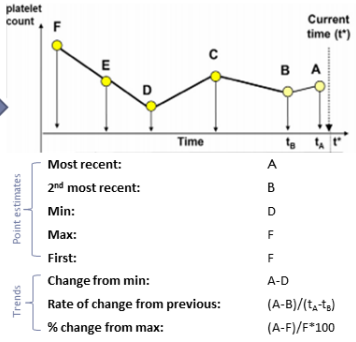
Study activities

- Background questionnaire
- Practice patient case review
- Patient case reviews (6 cases total)
  - Activities:
    - Review information, pretending as though preparing for rounds
    - Answer brief questionnaire
    - Give verbal presentation of patient case
  - Prediction tools options:
    - No risk prediction tool (2 cases)
    - Risk prediction tool with predictions only (2 cases)
    - Risk prediction tool with predictions and explanations (2 cases)
- Subjective assessment of prediction tools

In-hospital mortality risk model

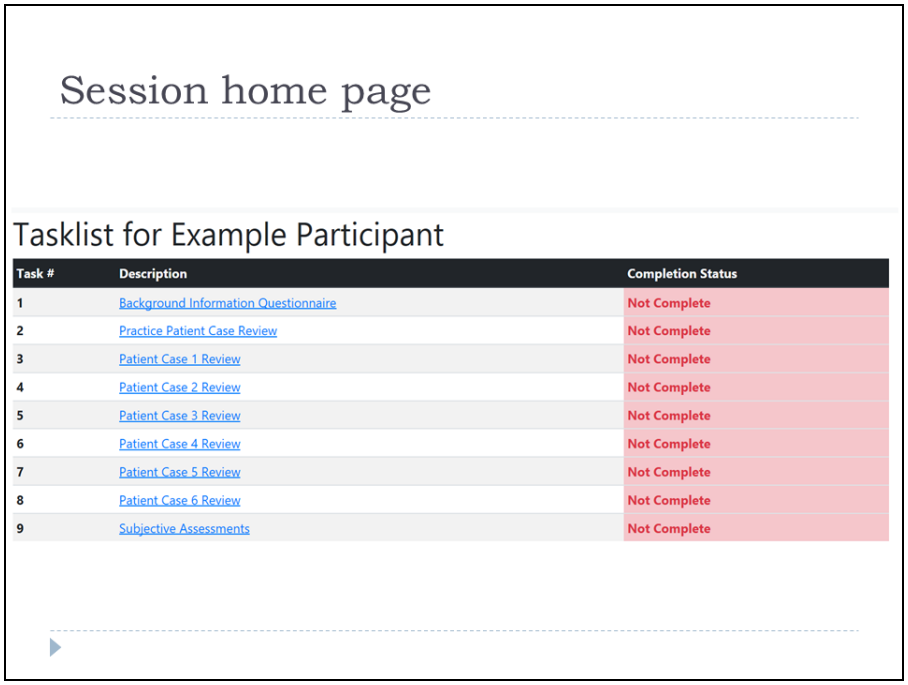
- Target
  - Risk of in-hospital death in next 24 hours for patients admitted to CHP PICU
- Predictors
  - Demographics
  - Diagnoses
  - Locations
  - Lab test results
  - Vital signs

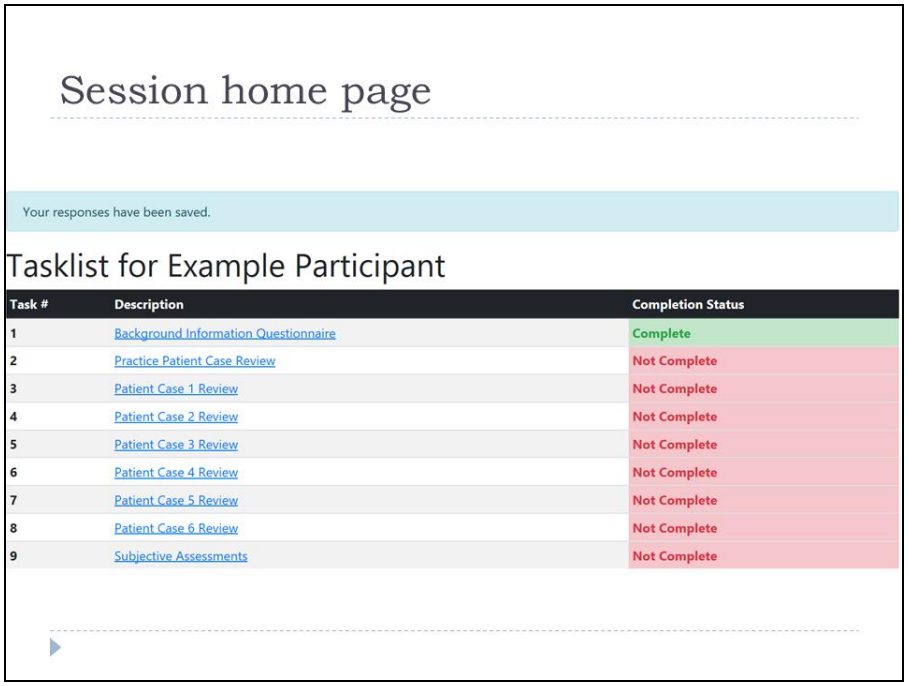
Summarized



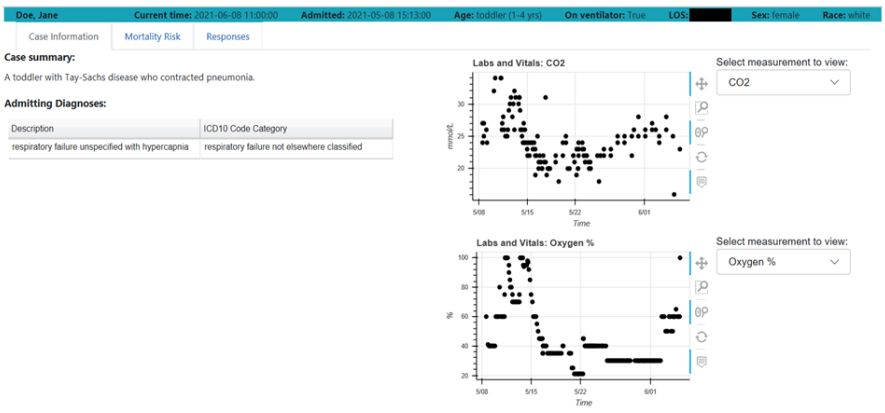
- Data
  - *Training:* data from PICU visits during 2015
  - *Testing:* data from PICU visits during 2016
- Performance:
  - Area under ROC curve: 0.93
  - Area under precision-recall curve: 0.78

Hauskrecht M, Batal I, Valko M, Visweswaran S, Cooper GF, Clermont G. Outlier detection for patient monitoring and alerting. *J Biomed Inform.* 2013;46(1):47-55.

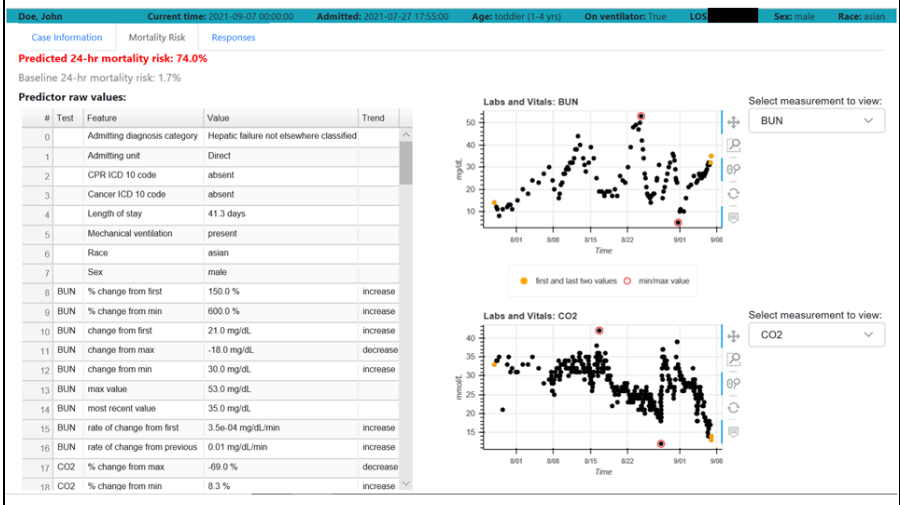

Session home page
Tasklist for Example Participant
Task # Description Completion Status
1 Background Information Questionnaire Not Complete
2 Practice Patient Case Review Not Complete
3 Patient Case 1 Review Not Complete
4 Patient Case 2 Review Not Complete
5 Patient Case 3 Review Not Complete
6 Patient Case 4 Review Not Complete
7 Patient Case 5 Review Not Complete
8 Patient Case 6 Review Not Complete
9 Subjective Assessments Not Complete


Session home page
Your responses have been saved.
Tasklist for Example Participant
Task # Description Completion Status
1 Background Information Questionnaire Complete
2 Practice Patient Case Review Not Complete
3 Patient Case 1 Review Not Complete
4 Patient Case 2 Review Not Complete
5 Patient Case 3 Review Not Complete
6 Patient Case 4 Review Not Complete
7 Patient Case 5 Review Not Complete
8 Patient Case 6 Review Not Complete
9 Subjective Assessments Not Complete

Patient case review

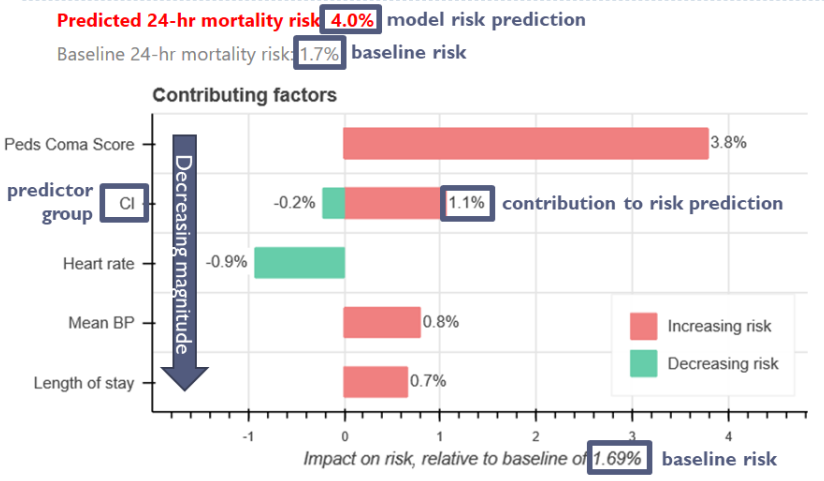


Prediction tool with predictions only

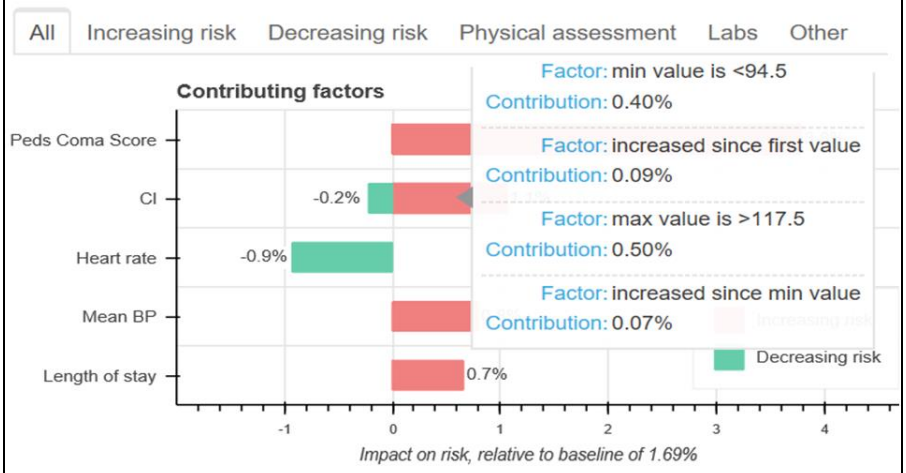


Prediction tool with explanations

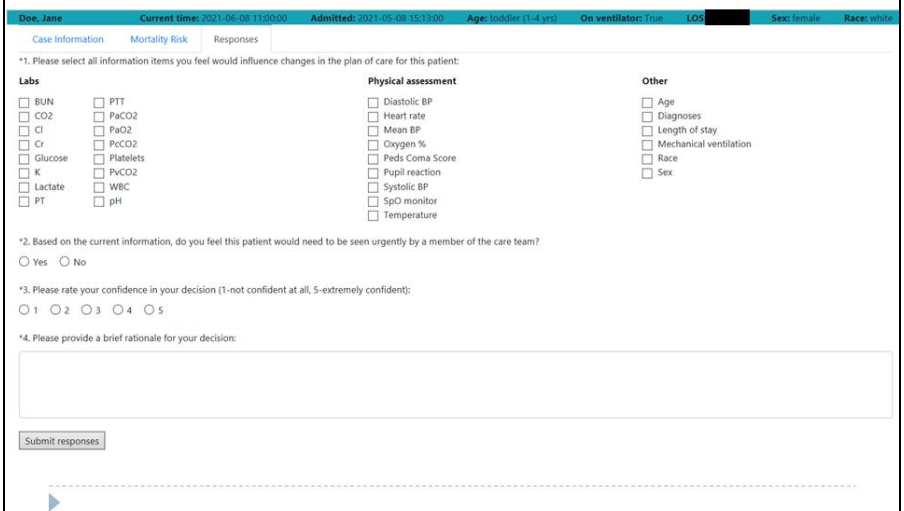
Prediction explanation cheat sheet



Prediction explanation cheat sheet



Patient case review



Patient case presentation

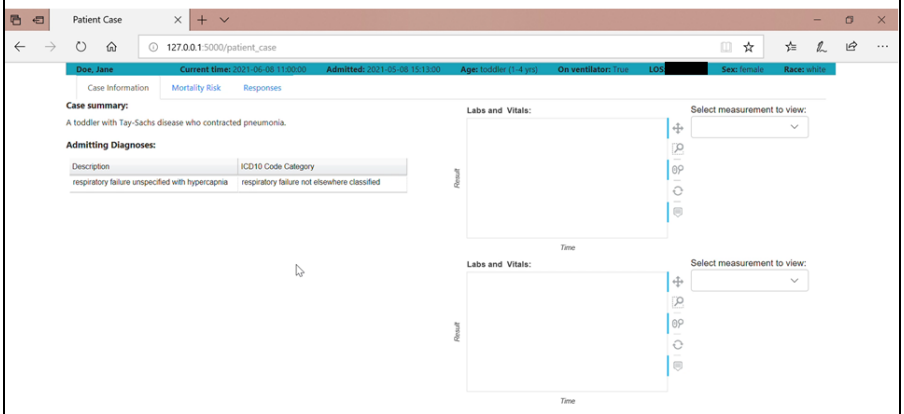
Your responses have been saved.

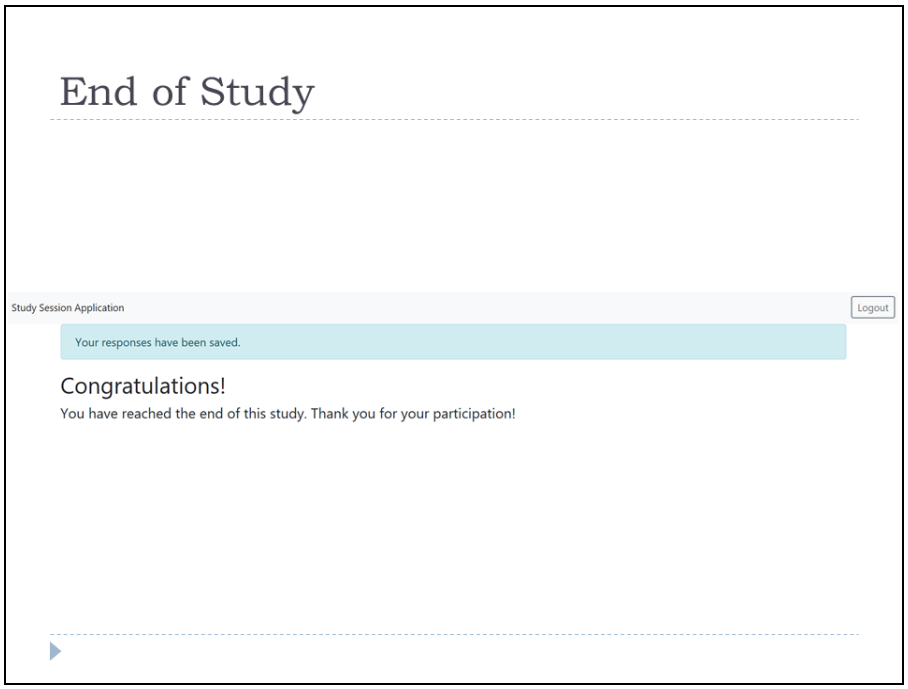
Patient Case Presentation

Please stop for a moment to verbally present your assessment of this patient, as you would present it to your team during rounds. The study moderator will start the recorder and let you know when to begin speaking. Once you have finished recording, please click the 'Done!' button below to continue with the study.

Done!

Interactive exploration




End of Study
Study Session Application
Logout
Your responses have been saved.
Congratulations!
You have reached the end of this study. Thank you for your participation!

Appendix E Evaluation Study Questionnaires

Background Questionnaire

*1. Please select the choice that most accurately describes your current clinical position:

- ○ Attending
- ○ Fellow
- ○ Resident
- ○ Nurse Practitioner
- ○ Physician Assistant
- ○ Other

*2. How long have you been in this position?

- ○ <1 year
- ○ 1 to <2 years
- ○ 2 to <3 years
- ○ >=3 years

Patient Case Questionnaire

*1. Please select all information items you feel would influence changes in the plan of care for this patient:

Labs

- ☐ BUN
- ☐ Cl
- ☐ CO2
- ☐ Cr
- ☐ Glucose
- ☐ K
- ☐ Lactate
- ☐ PaCO2
- ☐ PaO2
- ☐ PcCO2
- ☐ pH
- ☐ Platelets
- ☐ PT
- ☐ PTT
- ☐ PvCO2
- ☐ WBC

Physical assessment

- ☐ Diastolic BP
- ☐ Heart rate
- ☐ Mean BP
- ☐ Oxygen %
- ☐ Peds Coma Score
- ☐ Pupil reaction
- ☐ SpO monitor
- ☐ Systolic BP
- ☐ Temperature

Other

- ☐ Age
- ☐ Diagnoses
- ☐ Length of stay
- ☐ Mechanical ventilation
- ☐ Race
- ☐ Sex

*2. Based on the current information, do you feel this patient would need to be seen urgently by a member of the care team?

○ Yes ○ No

*3. Please rate your confidence in your decision for question 2 (1-not confident at all, 5-extremely confident):

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

*4. Please provide a brief rationale for your decision:

UTAUT Questionnaires

For the mortality risk model that provided **only predictions, *without* explanations**, please indicate your level of agreement with the following statements:

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
*1. I would find the system useful in my job.	○	○	○	○	○
*2. Using the system would increase my productivity.	○	○	○	○	○
*3. Using the system would make it easier to do my job.	○	○	○	○	○
*4. Using the system would enable me to accomplish tasks more quickly.	○	○	○	○	○
*5. My interaction with the system would be clear and understandable.	○	○	○	○	○
*6. I would find the system easy to use.	○	○	○	○	○
*7. It would be easy for me to become skillful at using the system.	○	○	○	○	○

For the mortality risk model that provided **predictions *with* explanations**, please indicate your level of agreement with the following statements:

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
*1. I would find the system useful in my job.	○	○	○	○	○
*2. Using the system would increase my productivity.	○	○	○	○	○
*3. Using the system would make it easier to do my job.	○	○	○	○	○
*4. Using the system would enable me to accomplish tasks more quickly.	○	○	○	○	○
*5. My interaction with the system would be clear and understandable.	○	○	○	○	○
*6. I would find the system easy to use.	○	○	○	○	○
*7. It would be easy for me to become skillful at using the system.	○	○	○	○	○

Bibliography

1. Cabitza F, Rasoini R, Gensini GF. Unintended Consequences of Machine Learning in Medicine. *JAMA*. 2017;318(6):517-518. doi:10.1001/jama.2017.7797
2. Bhatt U. Maintaining The Humanity of Our Models. November 2018. https://arxiv.org/pdf/1711.05791.pdf.
3. Feldman K, Davis D, Chawla N V. Scaling and contextualizing personalized healthcare: A case study of disease prediction algorithm integration. *J Biomed Inform*. 2015;57:377-385. doi:10.1016/j.jbi.2015.07.017
4. Beam AL, Kohane IS. Big Data and Machine Learning in Health Care. *JAMA*. 2018;319(13):1317-1318. doi:10.1001/jama.2017.18391
5. Shah ND, Steyerberg EW, Kent DM. Big Data and Predictive Analytics: Recalibrating Expectations. *JAMA*. 2018;320(1):27-28. doi:10.1001/jama.2018.5602
6. Vellido A. Societal Issues Concerning the Application of Artificial Intelligence in Medicine. *Kidney Dis (Basel, Switzerland)*. 2019;5(1):11-17. doi:10.1159/000492428
7. Deo RC. Machine Learning in Medicine. *Circulation*. 2015;132(20):1920-1930. doi:10.1161/CIRCULATIONAHA.115.001593
8. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform*. 2018;19(6):1236-1246. doi:10.1093/bib/bbx044
9. Nakamura F, Nakai M. Prediction Models - Why Are They Used or Not Used? *Circ J*. 2017;81(12):1766-1767. doi:10.1253/circj.CJ-17-1185
10. Katuwal GJ, Chen R. Machine Learning Model Interpretability for Precision Medicine. October 2016. http://arxiv.org/abs/1610.09045.
11. Bibal A, Frénay B. Interpretability of Machine Learning Models and Representations: an Introduction. In: *24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. Bruges, Belgium; 2016:77-82. https://www.elen.ucl.ac.be/Proceedings/esann/esannpdf/es2016-141.pdf.
12. Ahmad MA, Eckert C, Teredesai A. Interpretable Machine Learning in Healthcare. In: *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics - BCB '18*. New York, New York, USA: ACM Press; 2018:559-560. doi:10.1145/3233547.3233667
13. Cooper GF, Aliferis CF, Ambrosino R, et al. An evaluation of machine-learning methods for predicting pneumonia mortality. *Artif Intell Med*. 1997;9(2):107-138. doi:10.1016/s0933-3657(96)00367-3

14. Jovanovic M, Radovanovic S, Vukicevic M, Van Poucke S, Delibasic B. Building interpretable predictive models for pediatric hospital readmission using Tree-Lasso logistic regression. *Artif Intell Med*. 2016;72:12-21. doi:10.1016/j.artmed.2016.07.003

15. Hall P, Gill N. *An Introduction to Machine Learning Interpretability: An Applied Perspective on Fairness, Accountability, Transparency, and Explainable AI*. Sebastopol, CA, USA: O'Reilly Media, Inc.; 2018. http://www.oreilly.com/data/free/an-introduction-to-machine-learning-interpretability.csp.

16. Yang C, Delcher C, Shenkman E, Ranka S. Predicting 30-day all-cause readmissions from hospital inpatient discharge data. In: *2016 IEEE 18th International Conference on E-Health Networking, Applications and Services (Healthcom)*. IEEE; 2016:1-6. doi:10.1109/HealthCom.2016.7749452

17. Letham B, Rudin C, McCormick TH, Madigan D. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Ann Appl Stat*. 2015;9(3):1350-1371. doi:10.1214/15-AOAS848

18. Goodman B, Flaxman S. European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation." *AI Mag*. 2017;38(3):50-57. doi:10.1609/aimag.v38i3.2741

19. U.S. Food and Drug Administration. *Clinical and Patient Decision Support Software: Draft Guidance for Industry and Food and Drug Adminstration Staff*. Washington, D.C., USA; 2017. https://www.fda.gov/downloads/MedicalDevices/DeviceRegulationandGuidance/GuidanceDocuments/UCM587819.pdf.

20. Ras G, van Gerven M, Haselager P. Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges. In: Escalante HJ, Escalera S, Guyon I, et al., eds. *Explainable and Interpretable Models in Computer Vision and Machine Learning*. Springer, Cham; 2018:19-36. doi:10.1007/978-3-319-98131-4_2

21. Martens D, Vanthienen J, Verbeke W, Baesens B. Performance of classification models from a user perspective. *Decis Support Syst*. 2011;51(4):782-793. doi:10.1016/j.dss.2011.01.013

22. Pazzani MJ. Knowledge discovery from data? *IEEE Intell Syst their Appl*. 2000;15(2):10-12.

23. Johnson TL, Brewer D, Estacio R, et al. Augmenting Predictive Modeling Tools with Clinical Insights for Care Coordination Program Design and Implementation. *EGEMS (Washington, DC)*. 2015;3(1):1181. doi:10.13063/2327-9214.1181

24. Elish MC. The Stakes of Uncertainty: Developing and Integrating Machine Learning in Clinical Care. In: *Ethnographic Praxis in Industry Conference Proceedings*. Vol 2018. ; 2018:364-380. doi:10.1111/1559-8918.2018.01213

25. Barakat NH, Bradley AP, Barakat MNH. Intelligible support vector machines for diagnosis of diabetes mellitus. *IEEE Trans Inf Technol Biomed*. 2010;14(4):1114-1120. doi:10.1109/TITB.2009.2039485

26. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A Survey of Methods for Explaining Black Box Models. *ACM Comput Surv*. 2018;51(5):1-42. doi:10.1145/3236009

27. Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artif Intell*. 2019;267:1-38. doi:10.1016/j.artint.2018.07.007

28. Dosilovic FK, Brcic M, Hlupic N. Explainable artificial intelligence: A survey. In: *2018 41st International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. IEEE; 2018:0210-0215. doi:10.23919/MIPRO.2018.8400040

29. Freitas AA. Comprehensible classification models. *ACM SIGKDD Explor Newsl*. 2014;15(1):1-10. doi:10.1145/2594473.2594475

30. Lipton ZC. The Doctor Just Won't Accept That! In: *Proceedings of NIPS 2017 Symposium on Interpretable Machine Learning*. Long Beach, CA, USA; 2017. http://arxiv.org/abs/1711.08037.

31. Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M. Trends and Trajectories for Explainable, Accountable and Intelligible Systems. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. New York, New York, USA: ACM Press; 2018:1-18. doi:10.1145/3173574.3174156

32. Zhu J, Liapis A, Risi S, Bidarra R, Youngblood GM. Explainable AI for Designers: A Human-Centered Perspective on Mixed-Initiative Co-Creation. In: *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE; 2018:1-8. doi:10.1109/CIG.2018.8490433

33. Miller T, Howe P, Sonenberg L. Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences. In: *IJCAI 2017 Workshop on Explainable Artificial Intelligence*. Melbourne, Australia; 2017. http://arxiv.org/abs/1712.00547.

34. Raghupathi W, Raghupathi V. Big data analytics in healthcare: promise and potential. *Heal Inf Sci Syst*. 2014;2(1):3. doi:10.1186/2047-2501-2-3

35. Kilsdonk E, Peute LW, Jaspers MWM. Factors influencing implementation success of guideline-based clinical decision support systems: A systematic review and gaps analysis. *Int J Med Inform*. 2017;98:56-64. doi:10.1016/j.ijmedinf.2016.12.001

36. Kilsdonk E, Peute LWP, Knijnenburg SL, Jaspers MWM. Factors known to influence acceptance of clinical decision support systems. *Stud Health Technol Inform*. 2011;169:150-154. doi:10.3233/978-1-60750-806-9-150

37. Garg AX, Adhikari NKJ, McDonald H, et al. Effects of computerized clinical decision support systems on practitioner performance and patient outcomes: a systematic review. *JAMA*. 2005;293(10):1223-1238. doi:10.1001/jama.293.10.1223

38. Venkatesh V, Morris MG, Davis GB, Davis FD. User Acceptance of Information Technology: Toward a Unified View. *MIS Q*. 2003;27(3):425-478.

39. Venkatesh V, Sykes TA, Xiaojun Zhang. “Just What the Doctor Ordered”: A Revised UTAUT for EMR System Adoption and Use by Doctors. In: *2011 44th Hawaii International Conference on System Sciences*. IEEE; 2011:1-10. doi:10.1109/HICSS.2011.1

40. Wang D, Yang Q, Abdul A, Lim BY. Designing Theory-Driven User-Centric Explainable AI. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems - CHI ’19*. New York, New York, USA: ACM Press; 2019:1-15. doi:10.1145/3290605.3300831

41. Lim BY, Yang Q, Abdul A, Wang D. Why these Explanations? Selecting Intelligibility Types for Explanation Goals. In: *Joint Proceedings of the ACM IUI 2019 Workshops*. Los Angeles, CA, USA; 2019.

42. Ribera M, Lapedriza A. Can we do better explanations? A proposal of User-Centered Explainable AI. In: *Joint Proceedings of the ACM IUI 2019 Workshops*. Los Angeles, CA, USA; 2019.

43. Doshi-Velez F, Kim B. Towards A Rigorous Science of Interpretable Machine Learning. February 2017. http://arxiv.org/abs/1702.08608.

44. Mohseni S, Zarei N, Ragan ED. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. November 2018. http://arxiv.org/abs/1811.11839.

45. Kappen TH, van Loon K, Kappen MAM, et al. Barriers and facilitators perceived by physicians when using prediction models in practice. *J Clin Epidemiol*. 2016;70:136-145. doi:10.1016/j.jclinepi.2015.09.008

46. Wynants L, Collins GS, Van Calster B. Key steps and common pitfalls in developing and validating risk models. *BJOG*. 2017;124(3):423-432. doi:10.1111/1471-0528.14170

47. Lipton ZC. The Mythos of Model Interpretability. In: *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. New York, NY, USA; 2016. http://arxiv.org/abs/1606.03490.

48. Doran D, Schulz S, Besold TR. What Does Explainable AI Really Mean? A New Conceptualization of Perspectives. October 2017. http://arxiv.org/abs/1710.00794.

49. Karim A, Mishra A, Newton MH, Sattar A. Machine Learning Interpretability: A Science rather than a tool. July 2018. http://arxiv.org/abs/1807.06722.

50. Poursabzi-Sangdeh F, Goldstein DG, Hofman JM, Vaughan JW, Wallach H. Manipulating and Measuring Model Interpretability. February 2018. http://arxiv.org/abs/1802.07810.

51. Adadi A, Berrada M. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*. 2018;6:52138-52160. doi:10.1109/ACCESS.2018.2870052

52. Gilpin LH, Bau D, Yuan BZ, Bajwa A, Specter M, Kagal L. Explaining Explanations: An Overview of Interpretability of Machine Learning. May 2018. http://arxiv.org/abs/1806.00069.

53. Hoffman RR, Klein G. Explaining Explanation, Part 1: Theoretical Foundations. *IEEE Intell Syst*. 2017;32(3):68-73. doi:10.1109/MIS.2017.54

54. Krause J, Perer A, Bertini E. Using Visual Analytics to Interpret Predictive Machine Learning Models. In: *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. New York, NY, USA; 2016. http://arxiv.org/abs/1606.05685.

55. Zerilli J, Knott A, Maclaurin J, Gavaghan C. Transparency in Algorithmic and Human Decision-Making: Is There a Double Standard? *Philos Technol*. 2019;32(4):661-683. doi:10.1007/s13347-018-0330-6

56. Weld DS, Bansal G. The Challenge of Crafting Intelligible Intelligence. March 2018. http://arxiv.org/abs/1803.04263.

57. Herman B. The Promise and Peril of Human Evaluation for Model Interpretability. In: *NIPS 2017 Symposium on Interpretable Machine Learning*. Long Beach, CA, USA; 2017. http://arxiv.org/abs/1711.07414.

58. Ventocilla E, Helldin T, Riveiro M, Bae J. Towards a Taxonomy for Interpretable and Interactive Machine Learning. In: *XAI Workshop on Explainable Artificial Intelligence*. ; 2018:151-157. https://www.researchgate.net/publication/326979343.

59. Pieters W. Explanation and trust: what to tell the user in security and AI? *Ethics Inf Technol*. 2011;13(1):53-64. doi:10.1007/s10676-010-9253-3

60. Klein G. Explaining Explanation, Part 3: The Causal Landscape. *IEEE Intell Syst*. 2018;33(2):83-88. doi:10.1109/MIS.2018.022441353

61. Hoffman R, Miller T, Mueller ST, Klein G, Clancey WJ. Explaining Explanation, Part 4: A Deep Dive on Deep Nets. *IEEE Intell Syst*. 2018;33(3):87-95. doi:10.1109/MIS.2018.033001421

62. Molnar C. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable.*; 2018. https://christophm.github.io/interpretable-ml-book/.

63. Biran O, Cotton C. Explanation and Justification in Machine Learning : A Survey. In: *IJCAI-17 Workshop on Explainable Artificial Intelligence (XAI)*. Melbourne, Australia; 2017.

64. Ribeiro MT, Singh S, Guestrin C. Model-Agnostic Interpretability of Machine Learning. In: *2016 ICML Workshop on Human Interpretability in Machine Learning (WHI 2016)*. New York, NY, USA; 2016. http://arxiv.org/abs/1606.05386.

65. Hernandez PF. Lighting the black box: explaining individual predictions of machine learning algorithms. 2018.

66. Ribeiro MT, Singh S, Guestrin C. Nothing Else Matters: Model-Agnostic Explanations By Identifying Prediction Invariance. In: *NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems*. Barcelona, Spain; 2016. http://arxiv.org/abs/1611.05817.

67. Tintarev N, Masthoff J. Evaluating the effectiveness of explanations for recommender systems. *User Model User-adapt Interact*. 2012;22(4-5):399-439. doi:10.1007/s11257-011-9117-5

68. Allahyari H, Lavesson N. User-oriented assessment of classification model understandability. In: *11th Scandinavian Conference on Artificial Intelligence*. Trondheim, Norway; 2011.

69. Hoffman RR, Mueller ST, Klein G. Explaining Explanation, Part 2: Empirical Foundations. *IEEE Intell Syst*. 2017;32(4):78-86. doi:10.1109/MIS.2017.3121544

70. Grice HP. Logic and Conversation. In: *Syntax and Semantics 3: Speech Arts*. New York: Academic Press; 1975:41-58.

71. Samek W, Wiegand T, Müller K-R. Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models. August 2017. http://arxiv.org/abs/1708.08296.

72. Olah C, Satyanarayan A, Johnson I, et al. The Building Blocks of Interpretability. *Distill*. 2018;3(3). doi:10.23915/distill.00010

73. Shmueli G. To Explain or to Predict? *Stat Sci*. 2010;25(3):289-310. doi:10.1214/10-STS330

74. Krause J, Perer A, Ng K. Interacting with Predictions. In: *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16*. New York, New York, USA: ACM Press; 2016:5686-5697. doi:10.1145/2858036.2858529

75. Lundberg SM, Nair B, Vavilala MS, et al. Explainable machine learning predictions to help anesthesiologists prevent hypoxemia during surgery. *Nat Biomed Eng*. 2018;2(10):749-760. doi:10.1101/206540

76. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible Models for HealthCare. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*. New York, NY, USA: ACM Press; 2015:1721-1730. doi:10.1145/2783258.2788613

77. Choi E, Bahadori MT, Kulas JA, Schuetz A, Stewart WF, Sun J. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. In: *30th Conference on Neural Information Processing Systems (NIPS 2016)*. Barcelona, Spain; 2016. http://arxiv.org/abs/1608.05745.

78. Che Z, Purushotham S, Khemani R, Liu Y. Interpretable Deep Models for ICU Outcome Prediction. *AMIA . Annu Symp proceedings AMIA Symp*. 2016;2016:371-380. http://www.ncbi.nlm.nih.gov/pubmed/28269832.

79. Luo G. Automatically explaining machine learning prediction results: a demonstration on type 2 diabetes risk prediction. *Heal Inf Sci Syst*. 2016;4(2). doi:10.1186/s13755-016-0015-4

80. Van Belle VMCA, Van Calster B, Timmerman D, et al. A mathematical model for interpretable clinical decision support with applications in gynecology. *PLoS One*. 2012;7(3):e34312. doi:10.1371/journal.pone.0034312

81. Soininen H, Mattila J, Koikkalainen J, et al. Software tool for improved prediction of Alzheimer's disease. *Neurodegener Dis*. 2012;10(1-4):149-152. doi:10.1159/000332600

82. Kunapuli G, Varghese BA, Ganapathy P, et al. A Decision-Support Tool for Renal Mass Classification. *J Digit Imaging*. 2018;31(6):929-939. doi:10.1007/s10278-018-0100-0

83. Yang Y, Tresp V, Wunderle M, Fasching PA. Explaining Therapy Predictions with Layer-Wise Relevance Propagation in Neural Networks. In: *2018 IEEE International Conference on Healthcare Informatics (ICHI)*. IEEE; 2018:152-162. doi:10.1109/ICHI.2018.00025

84. Sha Y, Wang MD. Interpretable Predictions of Clinical Outcomes with An Attention-based Recurrent Neural Network. In: *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology,and Health Informatics - ACM-BCB '17*. New York, New York, USA: ACM Press; 2017:233-240. doi:10.1145/3107411.3107445

85. Valdes G, Luna JM, Eaton E, Simone CB, Ungar LH, Solberg TD. MediBoost: a Patient Stratification Tool for Interpretable Decision Making in the Era of Precision Medicine. *Sci Rep*. 2016;6(1):37854. doi:10.1038/srep37854

86. Liu N, Kumara S, Reich E. Explainable data-driven modeling of patient satisfaction survey data. In: *2017 IEEE International Conference on Big Data (Big Data)*. IEEE; 2017:3869-3876. doi:10.1109/BigData.2017.8258391

87. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. *J Am Med Informatics Assoc*. 2017;24(1):198-208. doi:10.1093/jamia/ocw042

88. Johnson AEW, Ghassemi MM, Nemati S, Niehaus KE, Clifton DA, Clifford GD. Machine Learning and Decision Support in Critical Care. *Proc IEEE Inst Electr Electron Eng*. 2016;104(2):444-466. doi:10.1109/JPROC.2015.2501978

89. Desai N, Gross J. Scoring systems in the critically ill: uses, cautions, and future directions. *BJA Educ*. 2019;19(7):212-218. doi:10.1016/j.bjae.2019.03.002

90. Lee J, Maslove DM, Dubin JA. Personalized Mortality Prediction Driven by Electronic Medical Data and a Patient Similarity Metric. Emmert-Streib F, ed. *PLoS One*. 2015;10(5):e0127428. doi:10.1371/journal.pone.0127428

91. Celi LA, Galvin S, Davidzon G, Lee J, Scott D, Mark R. A Database-driven Decision Support System: Customized Mortality Prediction. *J Pers Med*. 2012;2(4):138-148. doi:10.3390/jpm2040138

92. Awad A, Bader-El-Den M, McNicholas J, Briggs J. Early hospital mortality prediction of intensive care unit patients using an ensemble learning approach. *Int J Med Inform*. 2017;108(July):185-195. doi:10.1016/j.ijmedinf.2017.10.002

93. Tonekaboni S, Joshi S, McCradden MD, Goldenberg A. What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use. 2019. http://arxiv.org/abs/1905.05134.

94. Pollack AH, Tweedy CG, Blondon K, Pratt W. Knowledge crystallization and clinical priorities: evaluating how physicians collect and synthesize patient-related data. *AMIA . Annu Symp proceedings AMIA Symp*. 2014;2014:1874-1883. http://www.ncbi.nlm.nih.gov/pubmed/25954460.

95. Hallen SAM, Hootsmans NAM, Blaisdell L, Gutheil CM, Han PKJ. Physicians' perceptions of the value of prognostic models: the benefits and risks of prognostic confidence. *Health Expect*. 2015;18(6):2266-2277. doi:10.1111/hex.12196

96. Van Belle V, Van Calster B. Visualizing Risk Prediction Models. *PLoS One*. 2015;10(7):e0132614. doi:10.1371/journal.pone.0132614

97. Cabitza F, Zeitoun J-D. The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. *Ann Transl Med*. 2019;7(8):161-161. doi:10.21037/atm.2019.04.07

98. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. *Lancet (London, England)*. 1974;2(7872):81-84. doi:10.1016/s0140-6736(74)91639-0

99. Office of Management and Budget. Revisions to the standards for the classification of federal data on race and ethnicity. *Fed Regist*. 1997. https://nces.ed.gov/programs/handbook/data/pdf/Appendix_A.pdf.

100. Centers for Medicare & Medicaid Services. 2015 ICD-10-CM and GEMs. https://www.cms.gov/Medicare/Coding/ICD10/2015-ICD-10-CM-and-GEMs. Accessed January 5, 2019.

101. Fayyad UM, Irani KB. Multi-lnterval Discretization of Continuous-Valued Attributes for Classification learning. In: *13th International Joint Conference on Artificial Intelligence*. ; 1993:1022-1027.

102. Hall MA. Correlation-based Feature Selection for Machine Learning. 1999;(April). https://www.cs.waikato.ac.nz/~mhall/thesis.pdf.

103. Frank E, Hall MA, Witten IH. *The WEKA Workbench. Online Appendix for "Data Mining: Practical Machine Learning Tools and Techniques."* Fourth Edi. Hamilton, New Zealand: Morgan Kaufmann; 2016. https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf.

104. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA Data Mining Software: An Update. *SSIGKDD Explor*. 2009;11(1).

105. Reutemann P. python-weka-wrapper3. https://github.com/fracpete/python-weka-wrapper3.

106. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. *J Mach Learn Res*. 2011;12:2825-2830. http://dl.acm.org/citation.cfm?id=2078195%5Cnhttp://arxiv.org/abs/1201.0490.

107. Meyfroidt G, Güiza F, Ramon J, Bruynooghe M. Machine learning techniques to examine large patient databases. *Best Pract Res Clin Anaesthesiol*. 2009;23(1):127-143. doi:10.1016/j.bpa.2008.09.003

108. Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. In: *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*. New York, New York, USA: ACM Press; 2006:233-240. doi:10.1145/1143844.1143874

109. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2018. https://www.r-project.org/.

110. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1):77. doi:10.1186/1471-2105-12-77

111. Grau J, Grosse I, Keilwagen J. PRROC: computing and visualizing precision-recall and receiver operating characteristic curves in R. *Bioinformatics*. 2015;31(15):2595-2597. doi:10.1093/bioinformatics/btv153

112. Kennedy G, Gallego B. Clinical prediction rules: A systematic review of healthcare provider opinions and preferences. *Int J Med Inform*. 2019;123(November 2017):1-10. doi:10.1016/j.ijmedinf.2018.12.003

113. Edwards A. Explaining risks: turning numerical data into meaningful pictures. *BMJ*. 2002;324(7341):827-830. doi:10.1136/bmj.324.7341.827

114. Wadhwa R, Fridsma DB, Saul MI, et al. Analysis of a failed clinical decision support system for management of congestive heart failure. *AMIA . Annu Symp proceedings AMIA Symp*. November 2008:773-777. http://www.ncbi.nlm.nih.gov/pubmed/18999183.

115. Horsky J, Schiff GD, Johnston D, Mercincavage L, Bell D, Middleton B. Interface design principles for usable decision support: a targeted review of best practices for clinical prescribing interventions. *J Biomed Inform*. 2012;45(6):1202-1216. doi:10.1016/j.jbi.2012.09.002

116. Kannampallil TG, Jones LK, Patel VL, Buchman TG, Franklin A. Comparing the information seeking strategies of residents, nurse practitioners, and physician assistants in critical care settings. *J Am Med Informatics Assoc*. 2014;21(e2):e249-e256. doi:10.1136/amiajnl-2013-002615

117. Lee J, Maslove DM. Customization of a Severity of Illness Score Using Local Electronic Medical Record Data. *J Intensive Care Med*. 2017;32(1):38-47. doi:10.1177/0885066615585951

118. Pickering BW, Gajic O, Ahmed A, Herasevich V, Keegan MT. Data Utilization for Medical Decision Making at the Time of Patient Admission to ICU*. *Crit Care Med*. 2013;41(6):1502-1510. doi:10.1097/CCM.0b013e318287f0c0

119. Hall A, Walton G. Information overload within the health care system: a literature review. *Health Info Libr J*. 2004;21(2):102-108. doi:10.1111/j.1471-1842.2004.00506.x

120. Nushi B, Kamar E, Horvitz E. Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure. *Sixth AAAI Conf Hum Comput Crowdsourcing*. September 2018. http://arxiv.org/abs/1809.07424.

121. Kappen TH, Peelen LM. Prediction models. *Curr Opin Anaesthesiol*. 2016;29(6):717-726. doi:10.1097/ACO.0000000000000386

122. Pu P, Chen L. Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Syst*. 2007;20(6):542-556. doi:10.1016/j.knosys.2007.04.004

123. Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA, USA: ACM; 2016:1135-1144. http://arxiv.org/abs/1602.04938.

124. Lundberg S, Lee S-I. An unexpected unity among methods for interpreting model predictions. In: *NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems*. Barcelona, Spain; 2016. http://arxiv.org/abs/1611.07478.

125. Lundberg S, Lee S-I. A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems*. Long Beach, CA, USA; 2017:4765-4774. http://arxiv.org/abs/1705.07874.

126. Lundberg SM. SHAP (SHapley Additive exPlanations). https://github.com/slundberg/shap. Accessed October 10, 2018.

127. Bokeh Development Team. Bokeh: Python library for interactive visualization. http://www.bokeh.pydata.org.

128. Corbin J, Strauss A. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. 3rd ed. Los Angeles: SAGE Publications; 2008.

129. *NVivo Qualitative Data Analysis Software*. Version 12. QSR International Pty Ltd.; 2018.

130. The Pallets Projects. Flask: The Python micro framework for building web applications. https://github.com/pallets/flask.

131. WTForms. https://github.com/wtforms/wtforms.

132. SQLAlchemy. SQLAlchemy: The Python SQL Toolkit and Object Relational Mapper. https://github.com/sqlalchemy/sqlalchemy.

133. StataCorp. *Stata Statistical Software: Release 15*. College Station, TX: StataCorp LLC; 2017.

134. Hunter JD. Matplotlib: A 2D Graphics Environment. *Comput Sci Eng*. 2007;9(3):90-95. doi:10.1109/MCSE.2007.55

135. Matplotlib. https://github.com/matplotlib/matplotlib.

136. Jeffery AD, Novak LL, Kennedy B, Dietrich MS, Mion LC. Participatory design of probability-based decision support tools for in-hospital nurses. *J Am Med Informatics Assoc*. 2017;24(6):1102-1110. doi:10.1093/jamia/ocx060

137. Dekker FW, Ramspek CL, van Diepen M. Con: Most clinical risk scores are useless. *Nephrol Dial Transplant*. 2017;32(5):752-755. doi:10.1093/ndt/gfx073

138. Barr PJ, Elwyn G. Measurement challenges in shared decision making: putting the 'patient' in patient-reported measures. *Heal Expect*. 2016;19(5):993-1001. doi:10.1111/hex.12380

139. Li AC, Kannry JL, Kushniruk A, et al. Integrating usability testing and think-aloud protocol analysis with "near-live" clinical simulations in evaluating clinical decision support. *Int J Med Inform*. 2012;81(11):761-772. doi:10.1016/j.ijmedinf.2012.02.009

140. King AJ, Cooper GF, Clermont G, et al. Using machine learning to selectively highlight patient information. *J Biomed Inform*. 2019;100(August):103327. doi:10.1016/j.jbi.2019.103327

141. Ribeiro MT. Local Interpretable Model-Agnostic Explanations (lime). https://lime-ml.readthedocs.io/en/latest/. Accessed October 10, 2018.

142. Shapley LS. A value for n-person games. In: *Contributions to the Theory of Games*. Vol 28.; 1953:307-317.