VBridge: Connecting the Dots Between Features and Data to Explain Healthcare Models

Furui Cheng, Dongyu Liu, Fan Du, Yanna Lin, Alexandra Zytek, Haomin Li, Huamin Qu, and Kalyan Veeramachaneni

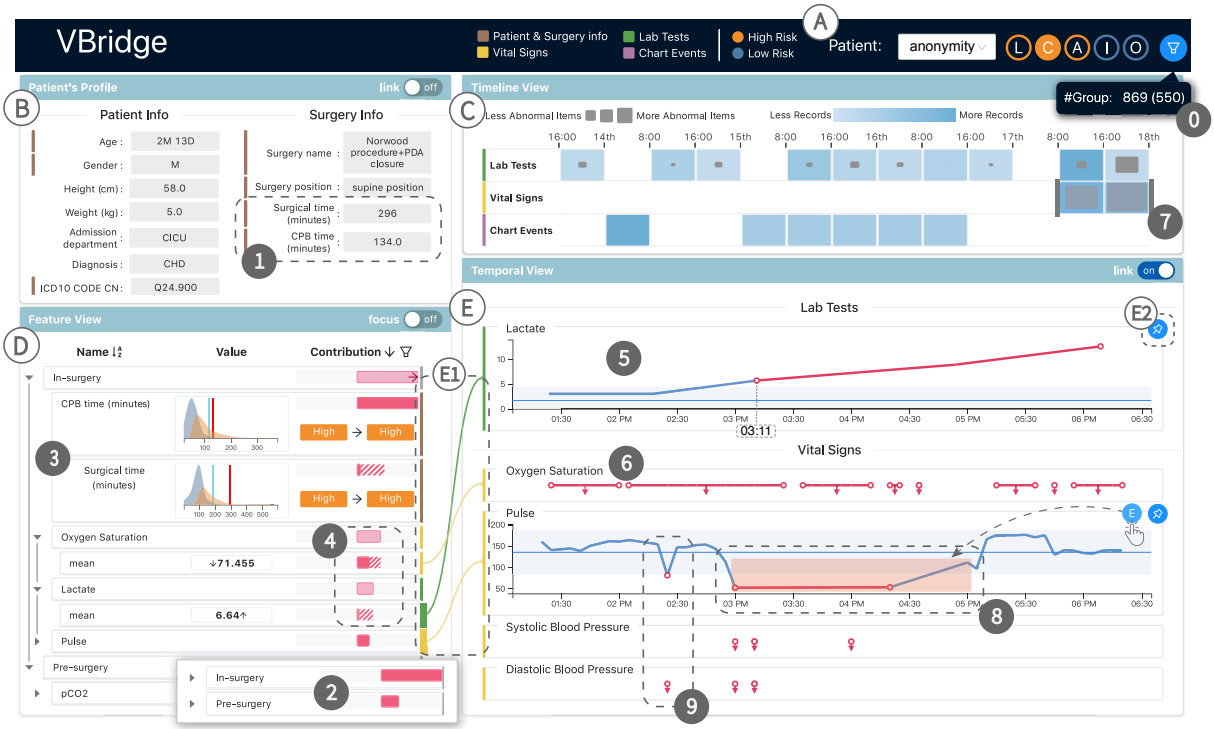

Fig. 1. The interface of VBridge facilitates clinicians' understanding and interpretation of ML model predictions. The header menu (A) allows clinicians to view prediction results, and to select a patient group for reference. The profile view (B) and the timeline view (C) show a summary of the target patient's health records. The feature view (D) shows feature-level explanations in a hierarchical display, linked to the temporal view (E) where healthcare time series are visualized to provide context for feature-level explanations. ⓪-⑨ show the step by step progression of a use case in which a clinician used the system to explore model explanations. After selecting a comparative group ⓪ and viewing the patient's profile ❶, he explored the feature-level explanations ❷-❹ to find the potential risk factors for the target patient. Then he referred to the patient's original records to gain an in-depth understanding ❺-⑨.

Abstract— Machine learning (ML) is increasingly applied to Electronic Health Records (EHRs) to solve clinical prediction tasks. Although many ML models perform promisingly, issues with model transparency and interpretability limit their adoption in clinical practice. Directly using existing explainable ML techniques in clinical settings can be challenging. Through literature surveys and collaborations with six clinicians with an average of 17 years of clinical experience, we identified three key challenges, including clinicians' unfamiliarity with ML features, lack of contextual information, and the need for cohort-level evidence. Following an iterative design process, we further designed and developed VBridge, a visual analytics tool that seamlessly incorporates ML explanations into clinicians' decision-making workflow. The system includes a novel hierarchical display of contribution-based feature explanations and enriched interactions that *connect the dots* between ML features, explanations, and data. We demonstrated the effectiveness of VBridge through two case studies and expert interviews with four clinicians, showing that visually associating model explanations with patients' situational records can help clinicians better interpret and use model predictions when making clinician decisions. We further derived a list of design implications for developing future explainable ML tools to support clinical decision-making.

Index Terms—Explainable Artificial Intelligence, Healthcare, Visual Analytics, Decision Making

1 Introduction

With the rapid proliferation of Electronic Health Records (EHRs), various prediction models based on machine learning (ML) techniques have been proposed for improving the quality of clinical care [46]. An

- *Furui Cheng, Yanna Lin, and Huamin Qu are with Hong Kong University of Science and Technology. E-mail: {fchengaa, ylindg, huamin}@ust.hk.*
- *Dongyu Liu, Alexandra Zytek, and Kalyan Veeramachaneni are with Massachusetts Institute of Technology. Dongyu Liu is the corresponding author. E-mail: {dongyu, zyteka, kalyanv}@mit.edu.*
- *Fan Du is with Adobe Research. E-mail: fdu@adobe.com.*
- *Haomin Li is with the Children's Hospital of Zhejiang University School of Medicine. Email: hmli@zju.edu.cn.*

Manuscript received 21 Mar. 2021; revised 13 June 2021; accepted 8 Aug. 2021. Date of publication 1 Oct. 2021; date of current version 22 Dec. 2021. Digital Object Identifier no. 10.1109/TVCG.2021.3114836

EHR stores an individual's health profile, from structured attributes like demographic information and medications, to unstructured ones, such as clinical notes and medical images. Prediction models, trained on patients' EHR data, can be useful for a wide range of medical outcomes [20, 47], including predicting a patient's *remaining length of stay*, the likelihood of *hospital readmission*, and *in-hospital mortality*.

Despite efforts by researchers and developers to improve the performance of these prediction models, challenges remain – including many associated with transparency and interpretability, which are particularly relevant in a highly regulated and risk-averse domain like healthcare [43, 51]. At the same time, XAI (eXplainable Artificial Intelligence) techniques and software tools continue to be developed, many of which have already proven powerful at elucidating the workings of "black-box" ML models. Nevertheless, prediction models built upon modern ML techniques have not yet been widely and reliably used in clinical decision support workflows [24, 40, 46]. By surveying related literature [1, 34, 43, 57, 58, 67] and working with 6 clinicians from a children's hospital, we found that the barriers preventing the application of XAI techniques in clinical settings are twofold.

First of all, clinicians engaging with XAI tools are often presumed to have sufficient technical expertise to understand and even improve ML models [1]. In reality, clinicians – who may have little to no technical background – are more likely to assess ML predictions through the lens of their domain expertise [51] rather than understand and improve the ML model from the technical point of view. This disconnection between the technology and its users is exacerbated by the fact that clinicians are rarely involved in discussions of explainability during the development of XAI tools [34]. As a result, the solutions provided by these tools are often intrinsically technical, leading to the difficulty for clinicians in understanding the "explanations" themselves [38, 57].

In addition, clinicians' workflows are often guided by individual patients and may require tailored explanations based on each patient's EHRs (*i.e.*, local explanations). Among the copious XAI approaches that support local explanations, feature contribution is one of the most popular. Approaches from this category illustrate the degree of contribution particular ML features make to a prediction outcome [43], which allows clinicians to directly compare model decisions with their own clinical judgment, especially when there is a disagreement. However, although these approaches have been extensively studied within the XAI field, there are still several significant challenges preventing their actual use in healthcare. Clinicians working with ML may run into problems in the following areas:

- **Understanding ML features.** Not every feature inputted to ML models is interpretable as-is by clinicians. For example, a patient's vital sign (*e.g.*, in-surgery heart rate) will be transformed into multiple ML features, each represented by an aggregate value (e.g., SD (standard deviation) or Trend (linear slope)) within a period (*a.k.a.*, feature engineering) [42]. While easily understood by an ML model, this form of representation is almost certainly unfamiliar and non-intuitive to clinicians – they may struggle to judge what, for example, a "high" Trend indicates, and the potential consequences.
- **Connecting to patients' original records.** Clinicians are more familiar with a given patient's original records than they are with ML features. In practice, they usually make decisions by referring to the raw data, such as laboratory test reports and vital signs from an anesthetic machine. However, feature contribution techniques only provide explanations on ML features, and do not deal with records directly [13, 16, 42, 55]. How to seamlessly connect these explanations to patients' original records remains an open question, and one that is underexplored.
- **Aligning with evidence.** Simply presenting a list of feature contributions in the form of numerical values does not allow clinicians to assess the trustworthiness of the model's predictions. Clinicians need to understand how feature contributions align with evidence-based medical practice [21, 57]. In this research, we propose using cohort-level statistics, available through hospital records, to provide this evidence. Clinicians can compare a target patient's feature values with reference values extracted from a cohort of similar patients.

The aforementioned challenges motivate us to design and develop a visual analytics solution that can seamlessly integrate feature-level explanations into a clinician's decision-making workflow. We followed a user-centric design process [58, 67] from the outset, working with 6 pediatric clinicians with an average of 17 years of work experience. We derived design requirements from a pilot study with these clinicians; then, by observing their interactions with our early-staged system, we summarized two workflows – forward analysis and backward analysis – preferred by clinicians with different levels of expertise. These requirements and workflows guided the overall design and development of VBridge, a **V**isualization system that **Bridge**s the gap between clinicians and ML models with tailored feature explanation algorithms and novel interaction and visualization techniques.

We adopted SHAP values [36] to generate contribution-based explanations of ML features and organized a large number of features in a hierarchy to facilitate interpretation. We developed a novel visualization – an interactive hierarchical feature list – to present such explanations to clinicians in a user-friendly manner and integrated tailored visual designs to allow clinicians to conduct reference-value-based analysis and what-if analysis at the feature level. To enable the connection between the feature explanations and the patient's raw records, we applied Deep Feature Synthesis [27] on EHR data to build traceable transformation paths between features and raw records. Based on this, we present a tailored algorithm to identify the most influential records for a given feature. The patient's original records are visualized in multiple coordinated views with different levels of detail. Various novel interactions, including linking and marking, help to visually associate the feature-level explanations and context information. The system was evaluated through two case studies and an expert interview with four clinicians, and results showed that our system is capable of supporting clinical decision-making.

To sum up, our contributions include:

- A summary of seven design requirements facilitating the interpretation of ML predictions to clinicians; and the identification of two workflows describing how they work with ML models with feature-level explanations and needed context information.
- A visual analytics system that integrates novel explanation algorithms and visualization and interaction techniques, to connect the dots between ML features, explanations, and health records for an improved clinicians' decision-making workflow.[1]
- Two case studies and an expert interview demonstrating the usefulness and efficiency of our system.

2 Related Work

2.1 Explainable Machine Learning in Clinical Predictions

We categorize existing XAI techniques in clinical research based on whether the provided interpretability is intrinsic or post-hoc.

Intrinsic interpretability. Models provide intrinsic interpretability by directly incorporating interpretability into their structures [12, 15, 23, 28, 32, 43]. Models in this category often use a simple structure to provide accurate and faithful explanations. For example, Kho *et al.* [28] used decision trees, which surface the set of rules driving the predictions, for predicting the genetic risk of type 2 diabetes. Despite their intrinsic interpretability, the performance of these models is bounded compared to advanced ML models (*e.g.*, deep neural networks), especially when handling complex clinical prediction tasks [20, 65]. Boosting and optimization techniques such as ensemble learning [23] can be used to enhance performance, but often at the cost of introducing additional complexity that impairs interpretability.

Recently, attention-based neural networks have begun to draw more focus [43]. Such models do not directly inform clinicians of the reasons behind a prediction, instead highlighting the portion of historical data (*e.g.*, clinical events) that have factored into it [12, 32]. Although deep learning models can produce accurate predictions, attention-based explanations may cause information overload and confuse clinicians due to the lack of clarity around how prediction results relate to

[1] The code is available at https://github.com/sibyl-dev/VBridge

the areas of attention [43]. It is also challenging for attention-based deep learning models to support multimodal learning while preserving good interpretability [65].

Post-hoc interpretability. Post-hoc methods take "black-box" ML models as inputs and then derive explanations for model predictions [10, 11, 35, 39, 48]. Unlike intrinsic interpretable models, post-hoc methods can be directly applied to existing models and thus are more flexible. One common approach is to use an intrinsically interpretable ML model to mimic a complex "black-box" ML model. For example, Che *et al.* [10] worked on acute lung injury (ALI) prediction and proposed a knowledge distillation method called mimic-learning, which uses gradient boosting trees to mimic the original deep learning model and provides rule explanations to clinicians.

Another type of work focuses on calculating feature contribution, which along with attention mechanism-based models, is considered to be one of the top popular approaches for supporting local explanations [43]. For example, Shapley Additive Explanations (SHAP) [35], which build on the Shapley value from cooperative game theory [54], have been applied to explain hypoxemia predictions and support early prevention during surgery [36].

Our work provides a post-hoc method for explaining existing models, in which we adopt feature-contribution-based XAI approaches. In particular, we use SHAP to compute how each ML feature contributes to a particular prediction. We present a tailored visualization technique to display feature contributions to clinicians in a scalable and user-friendly manner.

2.2 Electronic Health Records Visualization

We classify existing visualization techniques on EHR data [60] based on the criterion proposed by Rind *et al.* [49] – visualization for exploring health records from one patient or multiple patients.

Individual patient records. The goal of visualizing individual patient records is to provide individual patient summaries, as well as an efficient way to explore personal complex record data at different levels of detail. A patient's clinical records contain longitudinal data representing patient visits over time. One common way to summarize this history is through timeline-based visualizations, where events are placed on a horizontal timeline chronologically, using points or interval plots [45, 52, 53]. For events containing multiple attributes, glyphs [8] and additional tables [4, 17] are used to visually summarize events and facilitate more detailed explorations. To further improve scalability, researchers have explored aggregation-based methods [31] and substitution-based approaches [18] to show frequent patterns instead of event details. Another line of research focuses on event pattern searching, filtering, and grouping [64, 63], which supports fast and efficient data exploration.

In addition to discrete events, clinical signals collected during ICU or surgery are also commonly included in the EHR data. These are usually sampled at a higher frequency and can be viewed as continuous time series data. Xu *et al.* [66] used a spiral timeline to reveal periodic patterns of electrocardiogram data for arrhythmia detection. Our system builds on advances offered by prior visualization techniques to visualize a patient's EHR data at different levels of detail. We further tailored them for better interpreting ML predictions with feature-contribution-based XAI approaches.

Multiple patient records. A number of scenarios require the analysis of multiple patient records, from patient cohort monitoring to observational clinical research. A large number of works focus on visualizing longitudinal EHR data [5, 7, 18, 19, 25, 31, 32, 37, 44, 62], where glyphs [5, 7] and flow-based representations [18, 62] are often used for summarization. Other works focus on visualizing multivariate attributes or features transformed from the original records [2, 6, 30, 33, 41]. For example, Krause *et al.* [30] designed a glyph to visualize the quality of a feature under different metrics. In this work, we used aggregation-based methods to extract reference values from a cohort of patients. We proposed small intense, simple, and embeddable visualizations to show the reference values. These are integrated with the feature explanation view and raw record data visualization view to allow reference-value-based analysis.

3 Informing the Design

In this section, we introduce the pilot study and detail the design requirements and analysis workflow distilled from the study.

3.1 Pilot Study

The pilot study allowed us to understand how clinicians expect to use ML prediction models with feature contribution explanations to support their clinical decision-making. We followed the design study methodology from Sedlmair *et al.*'s work [50] and designed the pilot study as follows.

Participants: The study involved 6 clinicians (3 male and 3 female) from the Children's Hospital of Zhejiang University School of Medicine (ZJUCH): 2 chief physicians from the Cardiac Intensive Care Unit (CICU) (**P1-2**) and 4 residents from the Cardiology Department (**P3-6**). Among them, **P1-3** are more senior, with an average of 24.5 years of work experience (20, 29, and 24 years respectively), while the others (**P4-6**) have an average of 10.5 years of experience (13, 10, and 8.5 years respectively).

Presetting: The pilot study is based on a scenario of postoperative complication predictions. Patients may develop various complications after surgeries, some of which can be life-threatening. Predictions in an early phase can help clinicians identify high-risk patients and carefully choose postoperative caring plans. To support this scenario, we built a demo model on this prediction task. We worked with a biomedical data scientist (DS) from ZJUCH – a co-author of this paper, to carefully select a small set of features to train the model. We use SHAP values to show features' contributions to the prediction result.

Process: The study was divided into two sessions. We began the first session by performing one-on-one, semi-structured, hour-long interviews with all the participants. During the interview, the participants were presented with a low-fidelity mockup of our early system and taught some basic ML concepts. They were asked several questions about their understanding and concerns. Based on the feedback collected from this session, we formulated the initial design requirements. Over the next three months, we developed a high-fidelity prototype system, holding weekly meetings with **DS** to make sure our implementation continued to meet requirements.

In the second session, we presented the prototype system to three participants (**P2**, **P3**, **P5**) separately. They were asked to explore the system freely and completed several prediction tasks then, during which they were encouraged to think aloud to explain their thoughts. We observed, took notes, and collected their interaction processes. We then held an open discussion with them to further understand their behavior. The feedback collected from this round is further used to polish our design requirements and refine our system.

3.2 Design Requirements

We summarized seven design requirements and grouped them into: feature-level explorations (Feature), record-level explorations (Data), and explorations of feature-record connections (Bridge).

- **R1** Feature **Show features in a hierarchical structure.** All participants (**P1-6**) confirmed that it is challenging to explore hundreds of features extracted from diverse and heterogeneous sources. They all agreed with the idea of grouping relevant features semantically for a better exploration experience. For example, the aggregation values (*e.g.*, Mean, SD, and Trend) computed from the same series of data (*e.g.*, pulse) can be reasonably grouped.
- **R2** Feature **Provide features' reference values**. All participants (**P1-6**) agreed that the features, especially aggregate values that are unfamiliar to clinicians (*e.g.*, SD and Trend), should be presented alongside reference values, which describe the range of values that are considered normal. Because there are no existing reference values for most of the features, the system should calculate them using data from a relevant cohort.
- **R3** Feature **Provide flexible interactions** to support on-demand explorations. Participants follow different strategies when exploring features. Some participants (**P1**, **P3-5**) are only interested in the riskiest factors (*i.e.*, features with high positive con-

tributions), while some (**P2, P6**) were also interested in negatively contributing features that could be helpful for lowering the surgery risks in the future. Thus, the system should enable **sorting** and **filtering** to support different exploration paths. In addition, **P1** and **P2** expressed a further need to conduct what-if analyses on abnormal features to better understand their effects on predictions.

R4 Data **Provide an overview of the patient's records.** Patients have complex medical records, especially ICU patients, who may have substantially more data available than general patients. Good visualizations summarizing a patient's visiting history *"can save us a great amount of time in [familiarizing ourselves with] a patient's background"*, **P2** and **P6** confirmed.

R5 Data **Show record details with reference values.** Similar to **R2**, participants (**P1**, **P3-5**) suggested that showing patients' historical health record details along with reference values helps them to make more informed decisions. **P1** would like to know whether these values are within the 95% confidence interval (CI) of a statistical summary of similar patients.

R6 Bridge **Visually associate feature(s) with the patient's records.** All participants (**P1-6**) expressed the need to check the original records (*i.e.*, medical events) of particular features that interested them. *"Manually checking without any clues would take me 10-20 minutes"*, **P5** commented. Thus, visual associations of such correlations along with a tailored interaction mechanism should be enabled to support efficient back-and-forth analysis between features and their relevant original records.

R7 Bridge **Highlight temporal value patterns** that are influential to feature(s). Three participants (**P1**, **P3**, **P5**) expected the system to highlight high-risk time periods from long-lasting vital sign records related to the feature under investigation. **P1** showed particular interest in influential time periods containing a series of data points, rather than isolated anomalous data points which could be caused by errors.

3.3 Analysis Workflow

After analyzing the user interaction patterns and discussion notes from the second pilot study session, we summarized two general analysis workflows: **forward analysis** and **backward analysis**.

WF Forward analysis: clinicians inspect the data in an order similar to that of the direction of the data processing flow (*i.e.*, original records → features → predictions). They first make their own prediction based on the patients' original records, then compare their predictions with the model predictions, and finally check explanations to make decisions. Senior clinicians such as **P1** and **P2** preferred to start by viewing the patient's profile and forming initial hypotheses. They would then look directly at the original records (**R4**, **R5**) and check potentially influential observational records (*e.g.*, in-surgery lactate records). After making their own predictions based on this evidence, they seek confirmation from ML models and feature explanations (**R1-3**). If the model prediction and explanations agreed with their expectations – assigning high contribution values to the factors they thought were risky – their trust in the model was enhanced. If it didn't, they would refer to the patient's original records relevant to the features they were investigating to find more evidence-based on reference values (**R5-7**). They would either reject the model prediction, or would gain new knowledge based on this evidence.

WB Backward analysis: clinicians inspect the data in the opposite direction as the data processing flow. They check the model predictions and explanations first, and then trace back to the original records, finding evidence to support their decisions. When clinicians started without a clear diagnostic prediction, they began with a feature list with contribution explanations (**R1-3**). They then identified a set of features for further investigation. For static and familiar dynamic features (*e.g.*, in-surgery pulse), they compared the feature contributions with their expectations. For unfamiliar features, like SD or Trend of in-surgery systolic blood pressure, they preferred to check the details in the original records (**R5-7**). Sometimes these records were not sufficient to verify their hypothesis. They would then check the summary information of the original records (**R4**) in order to obtain different records from the same time, for correlation analysis.

4 Predictive Modeling

In this section, we first introduce the dataset, along with the prediction task we use as a running example for our research. Then we introduce how we extract ML features and generate explanations.

4.1 Data

VBridge takes structured EHR data collections as input. These are often organized as relational databases. In this work, we use the Paediatric Intensive Care (PIC) Database [69] as an example, which contains de-identified clinical data of paediatric patients admitted to ZJUCH. In particular, the dataset collects over 14,000 hospital admissions from 12,000 unique paediatric patients, aged 0–18 years, admitted to the critical care unit between 2010 and 2019. The PIC follows the same paradigm to store ICU patient clinical records as the widely-studied Medical Information Mart for Intensive Care (MIMIC-III) dataset [26], but puts more emphasis on paediatric patients. The dataset encompasses a number of information types, including demographics, surgery information, high-resolution vital sign measurements during surgery, laboratory test results, symptoms, medications, diagnostic codes, and mortality.

4.2 Running Example: Surgical Complication Prediction

To concretize our system's contributions, we utilize a running example – predicting complications after cardiac surgery – from a case study involving two clinicians from the ZJUCH team (**P1**, **P5**). This team was interested in using ML models to predict whether a patient is at risk of developing five types of complications after cardiac surgery: lung, cardiac, arrhythmia, infectious, and others, which are each annotated with their first letter (*i.e.*, L, C, A, I, O). A patient may experience multiple postoperative complications.

Working with the team and starting with the entire PIC dataset, we first selected 1,826 patients who underwent cardiopulmonary bypass-supported cardiac surgery. 456 (25.0%) of these patients developed postoperative complications. From the medical records of these patients, we mainly extracted three types of static features (demographics, surgery information, and diagnosis results) and three types of dynamic features whose values change over time (lab tests, surgery vital signs, and chart events[2]). In total, there were 1,724,805 lab test events, 450,989 chart events, and 754,213 data points from vital signs. In this example, our goal was to build 5 individual binary classifiers, each predicting one of the five complication types.

To be accepted by an ML model, a patient's raw medical data must be transformed into an ML-understandable format (*a.k.a.*, feature engineering) – namely, a feature vector (Fig. 2Ⓒ). Multiple feature vectors compose a feature matrix with each row describing one patient. Given the feature matrix and the target prediction column, in order to obtain the best ML model for our task, we applied Cardea [3] – an automated machine learning (AutoML) framework for EHR data. The framework evaluated 8 classifiers whose hyperparameters were optimized using AutoML for a higher performance score – the averaged AUC of 10 cross-validation folds (see Appendix). Finally, we obtained five models, each of which performed the best for one complication.

4.3 Feature Extraction

As shown in Fig. 2Ⓐ-left, EHR data from different sources is described as different entities, such as Admission (E_a), Lab Test (E_l), and Surgery (E_s). Entities are connected by reference keys (Fig. 2Ⓐ-right). In our running example of surgical complication prediction, we worked with our clinician collaborators and identified six feature types, both static and dynamic, with which to compose a patient's

[2]Chart events contain patients' routine vital signs (not during surgery) and additional information like inputs and outputs.

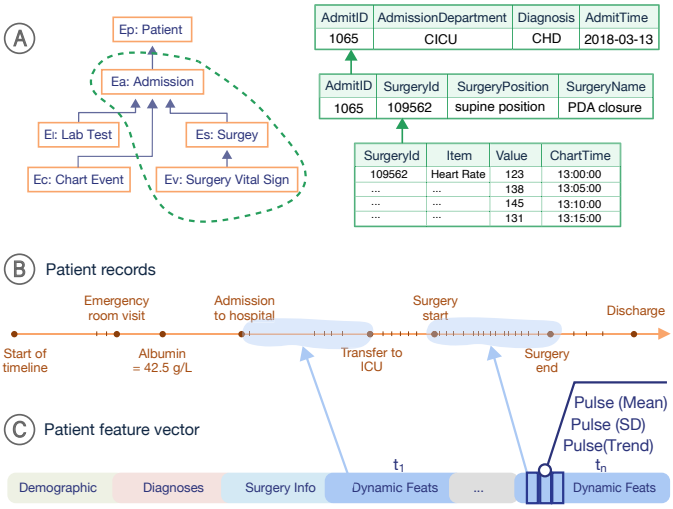

Fig. 2. Multiple-sourced EHR data is saved in different connected tables or entities (A). We use a DFS algorithm to extract it as a set of chronologically ordered patient records (B) and an ML feature vector (C).

feature vector. Our target patients are those who underwent cardiac surgery. To that end, we chose Surgery (E_s) as the target entity, and extracted the associated features including patient profile from E_p, surgery information from E_s, low-resolution time series (lab test and chart events) from E_l and E_c, and high-resolution time series from E_v.

Assembling patient feature vectors with DFS. We adapted Deep Feature Synthesis (DFS) [27] – an algorithm that automatically generates features out of relational tables – for our scenario to ensure that the connection between a patient's feature vector and raw records is traceable (**R6**). It works by following relationships between tables to a base field (*e.g.*, SurgeryId) and then sequentially applying transformation functions along the path to create the final feature. In the end, the algorithm will recursively extract all the associated features to our target entity (Surgery E_s) and each feature corresponds to a traceable path of length l between the final feature value and the raw record(s) of the source entity. The path is important for the purpose of visualization and the identification of influential records (**R6**, **R7**).

4.4 Feature Explanation

We applied SHAP values to provide feature-level explanations. However, for features that are unfamiliar to clinicians (*e.g.*, Trends), such explanations are not sufficient. Clinicians wish to further understand which time periods within the records (Fig. 2Ⓒ) are responsible for the feature of interest (**R6**, **R7**).

A common approach is occlusion sensitivity [68]. However, simply removing several medical records and observing how the prediction changes is not a feasible solution, because a surgical patient usually produces thousands of records – meaning that the model will not be sensitive if only a small number of records are removed. A similar approach, observing how relevant features change under the occlusion, has the same sensitivity issues. To solve the underlying sensitivity issues, we first calculate the influence of the records on the **relevant features' values** and identify the most influential time periods, using occlusion-based methods introduced below. Then we filter the influential record segments that push the relevant features' value away from the average level (*i.e.*, the reference value). Consider a scenario in which a patient's $Pulse(Trend)$, a major contributing feature, is significantly higher than the reference value. Clinicians may want to know during which period the records specifically cause a sudden increase in feature values, rather than all influential periods.

Computing record influence on a dynamic feature. Given a window size of k, a series of temporally ordered records $E = [x_1, x_2, ..., x_T]$, and a result array $\boldsymbol{v}$ of length T initialized to all 0s, we iteratively replace – or "occlude" – segments $E_{t:t+k}$ with some set of values, and increment t by after each step. We use window size k to reduce the impact of data quality issues and focus more on a segment of data (**R7**). We propose the use of a linear curve fit to the points in the window, which maintains smoothness while removing unique features within the window.

After each occlusion step, we recalculate the feature value, and store the change between the original feature value x and the updated feature value x' in the corresponding indices of $\boldsymbol{v}$: $\boldsymbol{v}_{t:t+k} = \boldsymbol{v}_{t:t+k} + (x - x')/\text{abs}(x)$. The results in $\boldsymbol{v}$ show the relative total influence of each record in E, based on how much and in which direction the feature changes when this point is removed. Notably, the real time computation of x' is possible because we store the traceable path between the relevant raw records and the feature value (Sec. 4.3).

Identifying the most influential time periods. Now that we have obtained an array of influence values $\boldsymbol{v}$, the next step is to highlight the most influential time periods (**R7**). This involves finding a threshold θ and identifying a list of segments $\boldsymbol{V}_{seq}$ with values above that threshold. Given that parametric approaches such as use of a Gaussian tail can be flawed when parametric assumptions are violated (*e.g.*, that the data follow Gaussian distributions), we adopted a non-parametric method without statistical assumptions. This method is adapted from the dynamic threshold computing method proposed by Hundman *et al.* [22]. We pick a threshold from the set: $\boldsymbol{\theta} = \mu(\boldsymbol{v}) + z\sigma(\boldsymbol{v})$, where $z \in \boldsymbol{z}$ is an ordered set of positive values indicating the number of SD (σ) above the mean (μ). The optimal θ is determined by:

$$\underset{\theta}{argmax} \quad \frac{\Delta\mu/\mu(\boldsymbol{v}) + \Delta\sigma/\sigma(\boldsymbol{v})}{|\boldsymbol{v}_a| + |\boldsymbol{V}_{seq}|^2}$$

where $\boldsymbol{v}_a = \{v_i \in \boldsymbol{v} | v_i > \theta\}$, $\Delta\mu = \mu(\boldsymbol{v}) - \mu(\boldsymbol{v} \setminus \boldsymbol{v}_a)$, $\Delta\sigma = \sigma(\boldsymbol{v}) - \sigma(\boldsymbol{v} \setminus \boldsymbol{v}_a)$, and $\boldsymbol{V}_{seq}$ = continuous sequences of $v_a \in \boldsymbol{v}_a$ The goal is to find a threshold θ that – once all values above it are eliminated – would lead to the maximum percent decrease in mean and SD of $\boldsymbol{v}$. Then we can obtain $\boldsymbol{V}_{seq}$ which represents the set of most "exceptionally" influential segments for a feature. We propose a novel visualization for showing this information to clinicians, detailed in Sec. 5.3.

5 VBridge

In this section, we continue with our running example – surgical complication prediction – to introduce our system designs.

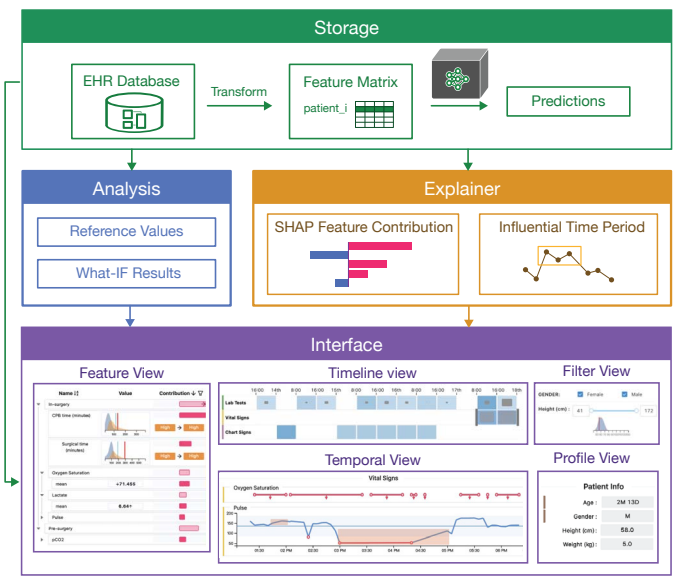

Fig. 3. The system architecture of VBridge, which consists of four tightly connected modules: a **storage** module, an **analysis** module, an **explainer** module, and an **interface** module.

5.1 System Overview

Fig. 3 illustrates the system architecture and the interactive analysis pipeline it supports. VBridge comprises four major modules: (1) storage, (2) analysis, (3) explainer, and (4) interface. The storage module saves all original patient records, a feature matrix with each row representing a patient's clinical features, and the ML prediction results. The analysis module supports dynamic calculation of reference values when a cohort of patients is selected, and real-time computation of the results for what-if analysis. The explainer module uses SHAP values to represent feature contributions, and identifies influential time periods for a given feature. Lastly, the interface module supplies multiple visual views, allowing a clinician to carry out his/her analysis using either a forward or backward workflow.

To show how the five views in the interface are connected, we assume that a clinician is using the backward analysis workflow to investigate the risk that a patient will have postoperative complications. First, he picks the patient and complication of interest from the top menu (Fig. 1Ⓐ). The five icons beside the selection box show the prediction results for the five types of complications – orange for positive and blue for negative. Next, he views the patient demographic, surgery, and admission information through the *Profile View*, identifies a cohort of patients as the reference patient group using the *Filter View*, and checks the patient number on the top menu (Fig. 1⓪).

Next, he begins formally investigating from the *Feature View*, which hierarchically shows ML-features, their contributions to the prediction result, and their reference values (**R1**, **R2**, **R3**). To further investigate the features of interest, he should check the original records of these features using the *Temporal View*; this view visualizes how a patient's clinical records change over time, along with the calculated reference range (**R5**) and the influential periods (**R7**). Multiple types of visual association, such as linking, filtering, and highlighting, are enabled to support the back-and-forth analysis between the *Feature View* and the *Temporal View* (**R6**). If the original records of the target feature (*e.g.*, Heart Rate) are not sufficient to verify the hypothesis, the clinician should then refer to the *Timeline View* to understand the overall situation and select contemporary medical records from other relevant features for further investigation (**R4**).

5.2 Feature View

The *Feature View* (Fig. 1Ⓓ) aims to allow clinicians to explore and understand the model's behavior at the feature level. To provide more consistent representations of these massive features, we have grouped relevant features according to suggestions from our clinician collaborators (**R1**). We first group the features (*e.g.*, Mean, SD, and Trend) that were extracted from the same series of medical events (*e.g.*, Pulse Records). We further divide the features (or groups) according to their temporal occurrence, which includes "pre-surgery" features (*e.g.*, demographics) and "in-surgery" features (*e.g.*, vital signs).

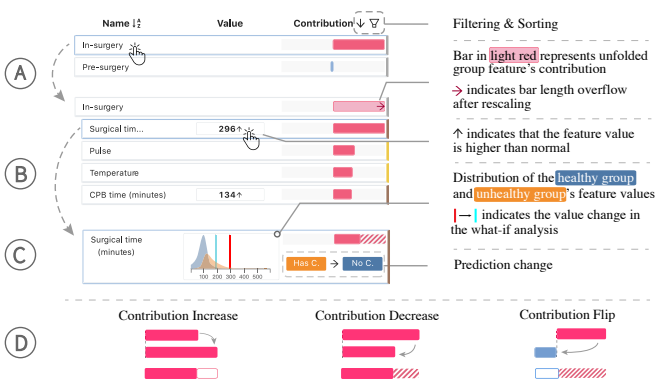

Fig. 4. Feature View. The design of the hierarchical feature list with different detail levels (A-C) and the design for visualizing contribution changes in the what-if analysis (D).

Hierarchical display (R1). We visualize the features in a hierarchical list where each row represents a feature or a feature group (Fig. 4). For each feature, we present the value and its contribution to the model prediction. We visually encode the contribution value with a horizontal bar, where the color encodes its sign (red for increasing complication risks and blue for decreasing risks) and the length encodes its magnitude (Fig. 4Ⓐ). For a feature group, we calculate group-level contributions by summing up the included features' additive contributions.

Based on the definition of Shapley values [54], group-level contributions can be explained as an approximation of the effects of removing this group of features from the model. Because clinicians have different goals or levels of knowledge, some expect to investigate the most fine-grained level of features (*e.g.* SD and Trend) while others may stop at the group level. The hierarchical feature list matches their demands well in this regard. Sorting and filtering by contributions are also supported to offer clinicians more control during explorations (**R1**).

References from cohorts (R2). In VBridge, reference values are calculated from a relevant cohort (*e.g.*, patients in the same age range) selected by users through the *Filter View*. The selected cohort is further divided into a low-risk group (*i.e.* no complications) and a high-risk group (*i.e.* one or more complications). We use the 95% CI of the low-risk group's mean value as the reference value range. We use an upward/downward arrow to indicate whether a value is beyond the upper/lower bound of the reference range (Fig. 4Ⓑ).

Clinicians can click the value area to inspect detailed value distributions of the low-risk group and the high-risk group (Fig. 4Ⓒ). For a continuous feature, the distribution is visualized with area charts, where a red line indicates the position of the feature value in relation to the target patient. For a categorical feature, we use bar charts to depict the distribution rather than area charts.

What-if analysis (R3). In evidence-based clinical practice, clinicians pay a lot of attention to anomalous records (*e.g.*, low Oxygen Saturation Rate) in the process of clinical reasoning. Our system marks values out of the reference range as anomalies. Clinicians are particularly interested in highly contributed features with anomalous values. For example in Fig. 4Ⓑ, the surgery time (296 minutes) is noted as an exceptionally high value by the upper arrow, and it also makes the highest contribution to the prediction. Our clinician collaborators had expressed strong interest in such cases, leading us to ask: If such a value is normal, does it still make a large contribution?

To answer this question, we designed a reference-value-based what-if analysis technique. Unlike open-ended what-if analysis techniques [61], we focus on one abnormal feature at a time, and make a minimal change to fit the reference range (*e.g.*, setting a high blood-pressure-related feature value to the upper bound of its reference range). Then we calculate and visualize changes in the prediction result and the target feature's contribution (Fig. 4Ⓒ - bottom right). We designed visualizations to encode the contribution changes while reserving the original contribution (*i.e.*, solid and dashed area) as context (Fig. 4Ⓓ). This approach provides clinicians with the most efficient and familiar way to verify their findings, especially when they aren't well-practiced in setting feature values.

5.3 Temporal View

The *Temporal View* visualizes a list of time series – each representing a type of time-varying clinical feature (*e.g.*., Heart Rate) – in order to provide context for feature-level explanations (Fig. 1Ⓔ). When clinicians find interesting features in the feature view (*e.g.*, Mean of Oxygen Saturation), they can append the corresponding time series records to the temporal view for further inspection (**R5**).

Each time series is visualized as a line chart (Fig. 5Ⓑ). We use the translucent blue area with a horizontal line in the middle to show the reference range (*i.e.*, 95% CI) and the mean value from the selected patient group. This design is familiar to clinicians and has frequently been used in clinical research [56]. In this paper's running example, children's observed values (*e.g.*, Pulse) vary significantly in different situations. In response, we compute reference values dynamically ac-

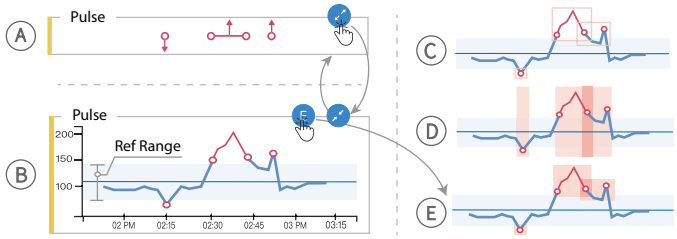

Fig. 5. Temporal View. The collapsed version (A) and the expanded version (B) of our time series record visualization, with anomalies highlighted. (C-E) show three design alternatives for highlighting influential time periods, where (E) is our last choice.

cording to the selected group of patients similar to the feature reference values (Sec. 5.2). We further empower the design to support the analysis of anomalies, concurrent patterns, and influential segments.

Highlighting anomalous records and enabling concurrent patterns analysis. Inside the line chart (Fig. 5Ⓑ), we use red dots and line segments to highlight out-of-reference-range records and time periods. To support clinicians inspecting multiple time series at the same time for concurrent pattern analysis (Fig. 1❾), we use a space-intensive design to only show the out-of-reference-range segments (Fig. 5Ⓐ). The arrow direction indicates whether a point (segment) is above or below the reference value range, whose design is consistent with the one used in the *Feature View* (Fig. 4Ⓑ).

Highlighting influential value patterns. Reference values provide clinicians with an evidence-based method for insight verification. However, clinicians are also curious to see how a ML model judges the influence of certain time periods, such as high-risk periods captured by the model (**R7**). We use the algorithm described in Sec. 4.4 to identify the most influential **non-overlapping** time segments $\boldsymbol{V}_{seq}$ and highlight them in the line chart (Fig. 5Ⓔ). However, multiple features (Mean, SD, Trend, etc.) may be associated with the same series of records (e.g, Pulse), so segments that are influential to different features can overlap. These overlapped areas often suggest highly influential time periods because they contribute to many features simultaneously. Inspired by Kim *et al.* [29], we consider three design alternatives (Fig. 5Ⓒ-Ⓔ) for highlighting prominent regions in the line chart. For Ⓒ, the bordered bounding box is accurate and clean, but not efficient at highlighting the overlapped area. For Ⓓ, the translucent full-height box highlights the overlap well, but is visually crowded. In accordance with our clinician collaborators, we finally chose the last design Ⓔ which combines the advantages of the other two designs.

5.4 Timeline View

The *Timeline View* (Fig. 1Ⓒ) provides an overview of the target patient's health records (**R4**). This view is the starting point for clinicians who use the forward analysis workflow (**WF**). In the meantime, it is also an indispensable part of the backward analysis workflow (**WB**), when a clinician desires to understand more contextual information about the patient. Through this view, clinicians can move additional medical records into the *Temporal View* for comparative analysis.

We use a matrix-based visualization [59] to show a summary of the target patient's medical events from different sources (lab tests, vital signs, and chart events) (Fig. 1Ⓒ). The horizontal timeline is divided into predefined, equal time intervals (*e.g.*, 1h, 4h, and 8h). Each cell contains the two pieces of information our clinician collaborators deemed most vital: (1) the background color encodes the number of events, with darker blue representing more events; and (2) the width of the inner box encodes the proportion of events containing out-of-reference-range values. For example, ▪ indicates that very few events occurred during this period and that most of them were normal, while ■ has the opposite meaning, and may call for an in-depth inspection. A similar design was used in Voila [9] to visualize the number of anomalous events of a region on a map.

Observing interesting cells in a particular row (*e.g.*, lab tests), clinicians can brush to select them, and click the "Go Temporal View" button to visualize all records from different items (*e.g.*, lab test items such as ALT, Glucose, and Lactate) in the *Temporal View* for a detailed investigation and comparative analysis.

5.5 Interaction

In addition to the basic interactions introduced above, VBridge offers two additional interactions, **linking** and **marking**, to facilitate better visual associations between features and their corresponding records.

Visually associating features and medical records (R6). Understanding connections between the feature elements (*i.e.*, rows in the feature list), and medical record elements (*e.g.*, temporal records in line charts and static information listed in the patient's profile) is not easy. Clinicians may need to scroll through a long list of features and compare names one-by-one. To make this easier, we propose the following novel and intuitive strategy. First, we use small colored bars to indicate the data source (*e.g.*, lab tests and vital signs) for both feature elements (on the right border) and medical record elements (on the left border). Then we draw curves to connect the associated feature elements with medical record elements (Fig. 1Ⓔ2). These curves are dynamically updated when users scroll down or join additional time-series records into the *Temporal View*.

Marking on medical records. To support the forward analysis workflow, clinicians are allowed to mark interesting medical record elements with "pins" (Fig. 1Ⓔ1). Associated feature elements are highlighted with a thicker bar. Clinicians can temporarily remove all other irrelevant features or feature groups by turning on the "focus" switch in the feature view's left-top corner.

6 Evaluation

In this section, we first introduce two case studies conducted with two clinicians (**P1**, **P5**) for evaluating whether VBridge and our proposed workflows (**WF**, **WB**) can support clinical decision-making. All clinicians also participated in the pilot study and development process, and are therefore familiar with the system.

6.1 Case Study I - Backward Analysis

We worked with **P5**, who has 10 years of work experience, to explore and the model's predictions about a two-month-old infant admitted to the CICU. The patient was predicted to be at high risk for various complications (L, C, A). The clinician was most interested in predicting cardiac complications (C), since they can lead to severe consequences. He first selected a group of patients in the same age range (*i.e.*, infants from 28 days to 12 months) to serve as references (**R2**, **R5**). This group included 869 patients (Fig. 1❿) of which 550 were healthy.

Exploring feature hierarchy (R1). The clinician first glanced at the patient's profile and noticed that two features ❶, surgery time and CPB (cardiopulmonary bypass) time, were much higher than usual. Keeping this in mind, he started exploring the feature view to check the features' contributions to the predicted complications. In the top level of the feature hierarchy ❷, he noticed that the contribution bar of the "in-surgery" feature group was much longer than that of the "pre-surgery" feature group, which means that the model mainly used information collected during surgery to make the prediction. The clinician then expanded the feature hierarchy and zoomed into a lower level to inspect the detailed explanations. Through sorting and filtering, he settled on a configuration where only the top 5 features or groups with the highest contributions were displayed in the list (**R1**, ❸). "*I like this control function and it helps me narrow down to a more focused display with only a few most important features*", he commented.

Understanding feature contributions (R2). He then noticed that the CPB time and the surgery time were the top 2 most important features whose values were both out of distribution ❸. He then commented "*This is exactly what I expected. Great to have a confirmation from the model about my previous suspicion*". He further wondered "*What would happen if their values go back to normal?*". We reminded him of the what-if analysis function (**R3**). Using this function,

he found no noticeable change in the prediction results for both features. However, he noticed that reducing the surgical time to the normal range decreased its contribution significantly. He thought "*The exceptionally long surgical time makes this feature positively contribute a lot to the model prediction, but other factors are still playing important roles because the prediction result does not change after what-if*".

The clinician then moved on to the two other features of interest – Oxygen Saturation and Lactate – as they are critical indicators of a patient's condition. Zooming into the most fine-grained feature level 4, he discovered that the contributions of these two features mostly came from the Mean features[3] which were either exceptionally low (Oxygen) or high (Lactate). He suspected such abnormal values should have considerable impacts on the model prediction and confirmed this suspicion after what-if analysis 4. He then showed further curiosity about the details of these abnormal values and commented "*I have to figure out when and why the lactate/oxygen saturation started to accumulate/drop. This is important for me to understand which catalysts, such as a patient's pre-existing condition or a surgeon's mistake, cause the results.*". So he selected the corresponding features to review them in the *Temporal View* (**R6**).

Inspecting features' influential records (R5). After the temporal view was displayed, he immediately observed that the Lactate level 5 was normal at the beginning of surgery, but started to increase after 2:00 PM and eventually went above the reference range after 3:00 PM. However, the Oxygen Saturation 6 was below the reference range during almost the entire surgical period (all the red downward-facing arrows). He commented "*I am so impressed by the smooth interaction and intuitive visualization design to guide me here. I think this patient might have cyanotic congenital heart disease, which could be the root cause for the hypoxemia and the lactate accumulation.*". He then decided to continue exploring the direct reason for the Lactate accumulation. He hypothesized that such accumulation was directly caused by the CPB process[4]. To confirm this, he referred to the timeline view and selected the vital signs during the surgery as references 7 (**R4**).

Taking a close look at the Pulse records, he noticed that the Pulse dropped to a very low level (50 BMP) at 3:00 PM and returned to normal at 5:00 PM 8. He confirmed this was the CPB period and explained "*During this time, the functions of the patient's heart and lungs were taken over by the CPB pump. That's why the patient's pulse looks abnormal.*" Comparing this period with the Lactate curve, he then rejected his earlier hypothesis, because the lactate had already reached a high level at 3:11 PM and in that case, the accumulation would have started earlier. Another interesting pattern – a sudden drop of Pulse around 2:30 PM 9 – caught his attention. He thought "*This was a rescue conducted at that time and is likely to be the key reason accounting for lactate accumulation*".

Noticing the sudden drop in Pulse, he was curious about whether the model "captured" this information while making predictions (**R7**). He then clicked the "explain" button to toggle the influential segments from the model's point of view 8. He noticed that most of the orange (influential) areas covered the CPB period. This fine-grained explanation is slightly different from his expectation – from his perspective, the model should also pay attention to the former sudden drop in Pulse. But in general, he agreed that the prediction was based on the most potentially critical medical records, and was trustworthy.

Summary. Through this exploration, **P5** was able to understand the most important features that led to the prediction, and to explore some interesting features and their corresponding records in depth. He decided to pay more attention to this patient, and considered using proactive treatments to avoid the situation getting worse in postoperative care.

6.2 Case Study II - Forward Analysis

We worked with **P1** – who has 20 years of experience in this field – to understand a prediction of high-risk lung-related complications made by the machine learning model.

Fig. 6. Case Study II – a use case involving understanding a prediction of high-risk lung-related complications following a **forward analysis workflow**. The clinician gained an overview of the patient's in-admission records from the timeline view 1. She then inspected the record details 2, marked interesting items 3, and formed the hypothesis. She verified the hypothesis with feature contributions 4 and influential time segments of the marked items 5, which were mostly within expectations. Finally, she explored features with unexpectedly high contributions 6-8, which helped her refine her judgements.

Gaining an overview of patient information (R4). The clinician started by checking the patient profile view. She thought everything (*e.g.*, surgical time and CPB time) was normal except for the patient's age (11 months), which was young for a VSD repair surgery. Then she looked at the timeline view and found the period during surgery (Fig. 6 1). In the row of lab tests, she noticed that most in-surgery test results were in normal ranges, indicated by the small grey inner rectangle. At the same time, vital signs had a slightly higher proportion of abnormal records. After the initial exploration, she found no solid evidence to indicate complication risks.

Inspecting record details (R5). She then checked the detailed lab tests and vital signs 2. She commented "*I don't find any big things. The three important indicators, Oxygen Saturation, Pulse, and Lactate, all look clean with no anomalous segments.*". She also noticed End-Tidal CO2 was below the reference range for a long period. Nevertheless, she hypothesized that the patient was not likely to have complications, which contradicted the model prediction. So she planned to refer to model explanations to figure out whether there were factors she had overlooked. She marked all four items 3 and continued to check the explanations in the feature view.

Comparing feature contributions with expectations. By tracing the links to the feature list (**R6**), she noticed that the feature group related to End-Tidal CO2 had a high positive contribution to the high-risk prediction 4. In contrast, features related to the other three items had slight negative contributions. She praised *"The explanation algorithm looks amazing. This actually matches what I expected. Now I am curious to see what the influential periods the model thinks to be"*. She clicked the "explain" button for help and then obtained the orange-highlighted area 5 which she thought was caused by CPB. The overlapped area with a deeper color also caught her attention, because multiple sub-features identified this area. She then said *"This is the critical changing point, but I might need more contextual information to test my thoughts"*.

She also noticed that Systolic Blood Pressure, Carboxyhemoglobin (COHb), and pre-surgery Red Cell Distribution Width (RDW) had the highest contributions. Among these, she noticed that the mean value

[3]Other features such as SD and Trend were filtered out due to their insignificant contributions.

[4]CPB is a technique that temporarily takes over the function of the heart and lungs during surgery, maintaining the circulation of blood and oxygen.

of COHb 6 and RDW 7 was higher than the reference range. She commented *"This is beyond my expectation. I know COHb is used to detect carbon monoxide (CO) toxicosis, but I never use this to judge whether a patient will develop complications"*. Through further inspection 8, she found that the COHb level was the highest right after the abnormal segment of End-Tidal CO2. She thought *"This might be unnoticeable factor in identifying the complications and I want to do further study with my team about it"*. As for the high RDW level, she realized that it might indicate that the patient suffered from iron-deficiency anemia, making them vulnerable to (lung) infections. This lab test does not tend to draw much attention from cardiac surgeons, so she had missed it earlier.

Summary. After the exploration, **P1** agreed that the patient was likely to have lung-related complications and decided to pay more attention to her. She was also curious about how COHb can be used to identify complications and considered studying it further.

7 Discussion

These case studies suggest that VBridge is helpful to clinicians and can support them in their decision-making. In addition to the case studies, we conducted semi-structured interviews with **P4** and **P6** by showing them the case study results and encouraging them to freely explore the system to collect additional feedback. We report and discuss feedback from all 4 clinicians as follows.

7.1 Design Implications

Feedback from these 4 clinicians led us to a set of important design considerations for all such projects, which we summarize as follows:

Applications of VBridge. All 4 participants generally commended the usefulness of VBridge in supporting diagnoses and expected to use the system to improve their daily workflow. **P1** expected to use the system to **make more accurate decisions**. She said *"Everyone sometimes may fall into blind spots. This tool can actually help me reduce the risk of making mistakes"*. **P4** expected to use the system to **communicate better** with other clinicians. He commented that *"A surgery involves collaborations between teams, …, people see data from different angles which might be biased somehow, …, I would trust VBridge and believe it can greatly facilitate the communication between teams"*. Both **P5** and **P6** suggested using VBridge to **help junior doctors** to make more accurate diagnoses.

Reference-value-based explanations. The reference values are vital in facilitating prediction interpretations for clinicians. Explanations like "the *Pulse.Mean*, whose value is below the reference range, has a high contribution to patient's cardiac complication" are easier for clinicians to understand and accept than purely reporting the contribution scores as confirmed by **P5**.

Feature hierarchy design. The hierarchical display of features was praised by all participants as it helps them avoid unnecessary details during exploration. Currently, there is no standard for designing the hierarchy of all healthcare features. However, ideas can be borrowed from the clinical forms used for communications between clinicians as suggested by **P1**.

Providing explanations with context. As demonstrated by the case studies, contextual information helps clinicians to understand explanations. Those in our study appreciated how the various visualization and interaction techniques in the system facilitated visual association between explanations and context. *"With the links, I can easily get connections between the features with their corresponding results,"*, as **P4** said. Also, **P1** suggested that "marking" is a very convenient interaction for checking explanations at will.

7.2 Limitations and Future Work

We introduce the limitations of our current work and future plans.

Feature interpretability. Our system only focuses on explaining predictions made from interpretable features (*i.e.*, features that have clear meanings and are extracted from a series of relevant health records). When the feature itself is hard for humans to understand (*e.g.*, features built from representation learning methods), the connections between features and health records can be very complex. In this case, the system will be less effective. An advanced method for tracing and storing such complex connections would be a good addition and remains to be explored.

Potential cognitive biases. Wang *et al.*'s work [58] suggests that a backward-oriented reasoning process (*i.e.*, first acquiring the diagnostic predictions, and then looking for supporting evidence) may lead to confirmation bias. Potential effects of cognitive biases on clinicians' decision-making when following different analysis workflows, and how our visualization designs may alleviate potential risks, have not been fully evaluated in this work. We plan to study this further by assessing the precision of clinicians' decisions when using VBridge.

Quality of EHR data. The poor quality of EHR data (*e.g.*, missing data) is a challenge to EHR data analysis in general. During the VBridge's development process, we also found many "False Positive" patterns caused by misrecorded data items (*e.g.*, a seeming cardiac arrest pattern was traced back to a faulty sensor). Currently, clinicians' prior knowledge is required to detect these data defects. In the future, we plan to investigate anomaly detection and visualization solutions to detect and encode any missing information in order to raise clinicians' awareness of missing data.

Precision of reference values. To improve the usability and precision of the dynamic reference value selection method, we plan to make the following extensions. First, we will automatically recommend relevant cohorts to clinicians for obtaining reference values. Second, we will derive time-varying reference values for temporal records (*e.g.*, Pulse), which are more applicable to surgical scenarios that are composed of multiple stages. Third, we will conduct experiments to understand the stability of the reference values (*i.e.*, how will the reference values change as the cohort changes over time?).

Visual scalability. Scalability issues occur in the temporal view when analyzing a signal with a large number of records. In addition, as the number of test items (rows) increases, finding interesting ones becomes less efficient, as more scrolling is required. In the future, we plan to scale up our approach by 1) segmenting long signals in different scales and 2) using searching and filtering techniques to facilitate the exploration of a vast number of complex signals.

Generalizability to other healthcare models. VBridge can be generalized to work on other prediction problems (*e.g.*, mortality predictions) and other ML models using the PIC dataset. However, adaptations (*e.g.*, formal descriptions of the entities and generated features) must be made to use VBridge with other EHR datasets (*e.g.*, MIMIC-III [26]), which is required by the feature extraction process (introduced in Sec. 4.3). In the future, we plan to improve generalizability by defining system inputs according to the Fast Healthcare Interoperability Resources (FHIR) standard [14], a general EHR data format.

8 Conclusion

In this work, we identified three key challenges limiting the use of ML in clinical settings, including clinicians' unfamiliarity with ML features, lack of contextual information, and the need for cohort-level evidence. We then introduced VBridge – a visual analytics system designed according to the requirements identified in a pilot study – to support clinicians using ML to make decisions with both forward and backward analysis workflows. We conducted two case studies and expert interviews with 4 clinicians. Their positive feedback and in-depth insights demonstrate the usefulness and effectiveness of the system. In particular, it reveals that visually associating model explanations with patients' situational records can help clinicians better interpret model predictions and use them to make clinical decisions.

Acknowledgments

We would like to thank Jianchuan Qi, Jie Jin, Lianglong Ma, Shanshan Shi, Xiuning Wu, and Xucong Shi from the Children's Hospital of Zhejiang University School of Medicine for their insightful feedback during the design process. We thank Arash Akhgari for his efforts in making the figures and feedback on the system interface designs. We also thank the anonymous reviewers for their valuable comments.

This research was supported in part by Hong Kong Theme-based Research Scheme grant T41-709/17N and a grant from MSRA.

REFERENCES

[1] M. A. Ahmad, C. Eckert, and A. Teredesai. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics*, pages 559–560, 2018.

[2] S. Alemzadeh, T. Hielscher, U. Niemann, L. Cibulski, T. Ittermann, H. Völzke, M. Spiliopoulou, and B. Preim. Subpopulation discovery and validation in epidemiological data. In *EuroVA@ EuroVis*, pages 43–47, 2017.

[3] S. Alnegheimish, N. Alrashed, F. Aleissa, S. Althobaiti, D. Liu, M. Alsaleh, and K. Veeramachaneni. Cardea: An open automated machine learning framework for electronic health records. In *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 536–545. IEEE, 2020.

[4] D. T. Bauer, S. A. Guerlain, and P. J. Brown. Evaluating the use of flowsheets in pediatric intensive care to inform design. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 50, pages 1054–1058. SAGE Publications Sage CA: Los Angeles, CA, 2006.

[5] J. Bernard, D. Sessler, J. Kohlhammer, and R. A. Ruddle. Using dashboard networks to visualize multiple patient histories: a design study on post-operative prostate cancer. *IEEE transactions on visualization and computer graphics*, 25(3):1615–1628, 2018.

[6] J. Bernard, D. Sessler, T. May, T. Schlomm, D. Pehrke, and J. Kohlhammer. A visual-interactive system for prostate cancer cohort analysis. *IEEE computer graphics and applications*, 35(3):44–55, 2015.

[7] H. S. G. Caballero, A. Corvò, P. M. Dixit, and M. A. Westenberg. Visual analytics for evaluating clinical pathways. In *2017 IEEE Workshop on Visual Analytics in Healthcare (VAHC)*, pages 39–46. IEEE, 2017.

[8] N. Cao, D. Gotz, J. Sun, and H. Qu. Dicon: Interactive visual analysis of multidimensional clusters. *IEEE transactions on visualization and computer graphics*, 17(12):2581–2590, 2011.

[9] N. Cao, C. Lin, Q. Zhu, Y.-R. Lin, X. Teng, and X. Wen. Voila: Visual anomaly detection and monitoring with streaming spatiotemporal data. *IEEE transactions on visualization and computer graphics*, 24(1):23–33, 2017.

[10] Z. Che, S. Purushotham, R. Khemani, and Y. Liu. Interpretable deep models for icu outcome prediction. In *AMIA Annual Symposium Proceedings*, volume 2016, page 371. American Medical Informatics Association, 2016.

[11] F. Cheng, Y. Ming, and H. Qu. Dece: Decision explorer with counterfactual explanations for machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1438–1447, 2020.

[12] E. Choi, M. T. Bahadori, J. A. Kulas, A. Schuetz, W. F. Stewart, and J. Sun. Retain: an interpretable predictive model for healthcare using reverse time attention mechanism. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 3512–3520, 2016.

[13] A. Eck, L. M. Zintgraf, E. de Groot, T. G. de Meij, T. S. Cohen, P. Savelkoul, M. Welling, and A. Budding. Interpretation of microbiota-based diagnostics by explaining individual classifier decisions. *BMC bioinformatics*, 18(1):1–13, 2017.

[14] FHIR. Fast healthcare interoperability resources. https://hl7.org/FHIR/, [Accessed: Apr 1st, 2021].

[15] A. Fleisher, B. Sowell, C. Taylor, A. Gamst, R. C. Petersen, L. Thal, et al. Clinical predictors of progression to alzheimer disease in amnestic mild cognitive impairment. *Neurology*, 68(19):1588–1595, 2007.

[16] W. Ge, J.-W. Huh, Y. R. Park, J.-H. Lee, Y.-H. Kim, and A. Turchin. An interpretable icu mortality prediction model based on logistic regression and recurrent neural networks with lstm units. In *AMIA Annual Symposium Proceedings*, volume 2018, page 460. American Medical Informatics Association, 2018.

[17] M. Ghassemi, M. Pushkarna, J. Wexler, J. Johnson, and P. Varghese. Clinicalvis: Supporting clinical task-focused design evaluation. *arXiv preprint arXiv:1810.05798*, 2018.

[18] D. Gotz and H. Stavropoulos. Decisionflow: Visual analytics for high-dimensional temporal event sequence data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1783–1792, 2014.

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

[20] H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and A. Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific data*, 6(1):1–18, 2019.

[21] R. B. Haynes, D. L. Sackett, W. S. Richardson, W. Rosenberg, and G. R. Langley. Evidence-based medicine: How to practice & teach ebm. *Canadian Medical Association. Journal*, 157(6):788, 1997.

[22] K. Hundman, V. Constantinou, C. Laporte, I. Colwell, and T. Soderstrom. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 387–395, 2018.

[23] A. Jalali and N. Pfeifer. Interpretable per case weighted ensemble method for cancer associations. *BMC genomics*, 17(1):1–10, 2016.

[24] P. B. Jensen, L. J. Jensen, and S. Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, 13(6):395–405, 2012.

[25] Z. Jin, S. Cui, S. Guo, D. Gotz, J. Sun, and N. Cao. Carepre: An intelligent clinical decision assistance system. *ACM Transactions on Computing for Healthcare*, 1(1):1–20, 2020.

[26] A. E. Johnson, T. J. Pollard, L. Shen, H. L. Li-Wei, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[27] J. M. Kanter and K. Veeramachaneni. Deep feature synthesis: Towards automating data science endeavors. In *2015 IEEE international conference on data science and advanced analytics (DSAA)*, pages 1–10. IEEE, 2015.

[28] A. N. Kho, M. G. Hayes, L. Rasmussen-Torvik, J. A. Pacheco, W. K. Thompson, L. L. Armstrong, J. C. Denny, P. L. Peissig, A. W. Miller, W.-Q. Wei, et al. Use of diverse electronic medical record systems to identify genetic risk for type 2 diabetes within a genome-wide association study. *Journal of the AMIA*, 19(2):212–218, 2012.

[29] D. H. Kim, V. Setlur, and M. Agrawala. Towards understanding how readers integrate charts and captions: A case study with line charts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2021.

[30] J. Krause, A. Perer, and E. Bertini. Infuse: Interactive feature selection for predictive modeling of high dimensional data. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1614–1623, 2014.

[31] M. Krstajic, E. Bertini, and D. Keim. Cloudlines: Compact display of event episodes in multiple time-series. *IEEE transactions on visualization and computer graphics*, 17(12):2432–2439, 2011.

[32] B. C. Kwon, M.-J. Choi, J. T. Kim, E. Choi, Y. B. Kim, S. Kwon, J. Sun, and J. Choo. Retainvis: Visual analytics with interpretable and interactive recurrent neural networks on electronic medical records. *IEEE transactions on visualization and computer graphics*, 25(1):299–309, 2018.

[33] B. C. Kwon, B. Eysenbach, J. Verma, K. Ng, C. De Filippi, W. F. Stewart, and A. Perer. Clustervision: Visual supervision of unsupervised clustering. *IEEE transactions on visualization and computer graphics*, 24(1):142–151, 2017.

[34] Z. C. Lipton. The doctor just won't accept that! *arXiv preprint arXiv:1711.08037*, 2017.

[35] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017.

[36] S. M. Lundberg, B. Nair, M. S. Vavilala, M. Horibe, M. J. Eisses, T. Adams, D. E. Liston, D. K.-W. Low, S.-F. Newman, J. Kim, et al. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10):749, 2018.

[37] S. Malik, F. Du, M. Monroe, E. Onukwugha, C. Plaisant, and B. Shneiderman. Cohort comparison of event sequences with balanced integration of visual analytics and statistics. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 38–49, 2015.

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

[39] Y. Ming, H. Qu, and E. Bertini. Rulematrix: Visualizing and understanding classifiers with rules. *IEEE transactions on visualization and computer graphics*, 25(1):342–352, 2018.

[40] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific reports*, 6(1):1–10, 2016.

[41] J. Müller, M. Stoehr, A. Oeser, J. Gaebel, M. Streit, A. Dietz, and S. Oeltze-Jafra. A visual approach to explainable computerized clinical decision support. *Computers & Graphics*, 91:1–11, 2020.

[42] S. Nemati, A. Holder, F. Razmi, M. D. Stanley, G. D. Clifford, and T. G. Buchman. An interpretable machine learning model for accurate prediction of sepsis in the icu. *Critical care medicine*, 46(4):547, 2018.

[43] S. N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J. H.

Chen, X. Liu, and Z. He. Explainable artificial intelligence models using real-world electronic health record data: a systematic scoping review. *Journal of the AMIA*, 27(7):1173–1185, 2020.

[44] D. Phan, A. Paepcke, and T. Winograd. Progressive multiples for communication-minded visualization. In *Proceedings of Graphics Interface 2007*, pages 225–232, 2007.

[45] C. Plaisant, R. Mushlin, A. Snyder, J. Li, D. Heller, and B. Shneiderman. Lifelines: using visualization to enhance navigation and analysis of patient records. In *The craft of information visualization*, pages 308–312. Elsevier, 2003.

[46] A. Rajkomar, J. Dean, and I. Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[47] A. Rajkomar, E. Oren, K. Chen, A. M. Dai, N. Hajaj, M. Hardt, P. J. Liu, X. Liu, J. Marcus, M. Sun, et al. Scalable and accurate deep learning with electronic health records. *NPJ Digital Medicine*, 1(1):1–10, 2018.

[48] M. T. Ribeiro, S. Singh, and C. Guestrin. " 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*, pages 1135–1144, 2016.

[49] A. Rind, T. D. Wang, W. Aigner, S. Miksch, K. Wongsuphasawat, C. Plaisant, and B. Shneiderman. Interactive information visualization to explore and query electronic health records. *Foundations and Trends in Human-Computer Interaction*, 5(3):207–298, 2013.

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

[51] M. Sendak, M. C. Elish, M. Gao, J. Futoma, W. Ratliff, M. Nichols, A. Bedoya, S. Balu, and C. O'Brien. "the human body is a black box" supporting clinical decision-making with deep learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 99–109, 2020.

[52] Y. Shahar and C. Cheng. Intelligent visualization and exploration of time-oriented clinical data. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences. 1999. HICSS-32. Abstracts and CD-ROM of Full Papers*, pages 12–pp. IEEE, 1999.

[53] Y. Shahar, D. Goren-Bar, D. Boaz, and G. Tahan. Distributed, intelligent, interactive visualization and exploration of time-oriented clinical data and their abstractions. *Artificial intelligence in medicine*, 38(2):115–135, 2006.

[54] L. S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[55] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pages 3145–3153. PMLR, 2017.

[56] B. Stubbs, D. C. Kale, and A. Das. Simtwentyfive: An interactive visualization system for data-driven decision support. In *AMIA Annual Symposium Proceedings*, volume 2012, page 891. American Medical Informatics Association, 2012.

[57] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine Learning for Healthcare Conference*, pages 359–380. PMLR, 2019.

[58] D. Wang, Q. Yang, A. Abdul, and B. Y. Lim. Designing theory-driven user-centric explainable ai. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–15, 2019.

[59] D. Weng, C. Zheng, Z. Deng, M. Ma, J. Bao, Y. Zheng, M. Xu, and Y. Wu. Towards better bus networks: A visual analytics approach. *IEEE transactions on visualization and computer graphics*, 27(2):817–827, 2021.

[60] V. L. West, D. Borland, and W. E. Hammond. Innovative information visualization of electronic health record data: a systematic review. *Journal of the AMIA*, 22(2):330–339, 2015.

[61] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE transactions on visualization and computer graphics*, 26(1):56–65, 2019.

[62] K. Wongsuphasawat and D. Gotz. Outflow: Visualizing patient flow by symptoms and outcome. In *IEEE VisWeek Workshop on Visual Analytics in Healthcare, Providence, Rhode Island, USA*, pages 25–28. American Medical Informatics Association, 2011.

[63] K. Wongsuphasawat, C. Plaisant, M. Taieb-Maimon, and B. Shneiderman. Querying event sequences by exact match or similarity search: Design and empirical evaluation. *Interacting with computers*, 24(2):55–68, 2012.

[64] K. Wongsuphasawat and B. Shneiderman. Finding comparable temporal categorical records: A similarity measure with an interactive visualization. In *2009 IEEE Symposium on Visual Analytics Science and Technology*, pages 27–34. IEEE, 2009.

[65] C. Xiao, E. Choi, and J. Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the AMIA*, 25(10):1419–1428, 2018.

[66] K. Xu, S. Guo, N. Cao, D. Gotz, A. Xu, H. Qu, Z. Yao, and Y. Chen. Ecglens: Interactive visual exploration of large scale ecg data for arrhythmia detection. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–12, 2018.

[67] Q. Yang, A. Steinfeld, and J. Zimmerman. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2019.

[68] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014.

[69] X. Zeng, G. Yu, Y. Lu, L. Tan, X. Wu, S. Shi, H. Duan, Q. Shu, and H. Li. Pic, a paediatric-specific intensive care database. *Scientific data*, 7(1):1–8, 2020.