ClinicalVis: Supporting Clinical Task-Focused Design Evaluation

A Visualization-Based Prototype System to Explore the Interaction of Healthcare Providers and Electronic Healthcare Records

Marzyeh Ghassemi
Verily, University of Toronto
Toronto, Ontario
marzyeh@cs.toronto.edu

Mahima Pushkarna
Google Brain
Cambridge, Massachusetts
mahima@google.com

James Wexler
Google Brain
Cambridge, Massachusetts
jwexler@google.com

Jesse Johnson
Verily, Sanofi
Cambridge, Massachusetts

Paul Varghese
Verily
Cambridge, Massachusetts
paulvarghese@verily.com

ABSTRACT

Making decisions about what clinical tasks to prepare for is multi-factored, and especially challenging in intensive care environments where resources must be balanced with patient needs. Electronic health records (EHRs) are a rich data source, but are task-agnostic and can be difficult to use as summarizations of patient needs for a specific task, such as "could this patient need a ventilator tomorrow?" In this paper, we introduce *ClinicalVis*, an open-source EHR visualization-based prototype system for task-focused design evaluation of interactions between healthcare providers (HCPs) and EHRs. We situate ClinicalVis in a task-focused proof-of-concept design study targeting these interactions with real patient data. We conduct an empirical study of 14 HCPs, and discuss our findings on usability, accuracy, preference, and confidence in treatment decisions. We also present design implications that our findings suggest for future EHR interfaces, the presentation of clinical data for task-based planning, and evaluating task-focused HCP/EHR interactions in practice.

KEYWORDS

Data-enabled Design; Healthcare Providers; Health; Data; Exploration; Clinical Informatics.

Under Review, ,

ACM ISBN ..
https://doi.org/.

ACM Reference Format:
Marzyeh Ghassemi, Mahima Pushkarna, James Wexler, Jesse Johnson, and Paul Varghese. 2019. ClinicalVis: Supporting Clinical Task-Focused Design Evaluation: A Visualization-Based Prototype System to Explore the Interaction of Healthcare Providers and Electronic Healthcare Records. In *Proceedings of (Under Review)*. ACM, New York, NY, USA, 15 pages. https://doi.org/.

1 INTRODUCTION

Patient Electronic Health Records (EHR) contain a wealth of heterogeneous data, leading to exciting opportunities in both information visualization and clinical support. Planning for safe and effective clinical care depends on the ability to parse this heterogeneous clinical data, and derive an understanding of a patient's health state. Data visualization techniques are known to improve rich data pattern communication and reduce overall cognitive load [59], which in turn can help Health Care Providers (HCPs) efficiently extract accurate information [8]. However, commercially available EHRs are often cognitively cumbersome to use, and EHR usability is a well-established HCP pain point [32].

Commercial EHRs are task-agnostic, support linear (rather than dynamic) care coordination processes, and are optimized for billing rather than HCP and patient use [5, 51]. They are also closed-source, making them difficult to evaluate [36]. Even taking screen images of a leading EHR provider can be viewed as legally prohibited [35, 64]. Access to clinical data — and clinical environments — is often limited for privacy reasons, leading professionals tasked with improving EHR designs to work with secondary information rather than real-world use cases [36].

There is currently no open-source framework that provides a visual summary of patient information to HCPs for planning specific clinical task, and subsequently evaluates

HCP response. To support the evaluation of HCPs-EHR interaction for task-specific clinical planning, our multidisciplinary team assessed a need to (a) select realistic clinical tasks HPCs need to plan for, (b) create a prototype system to view EHRs on, (c) identify a set of real patient cases where received care was recorded for each task, and (d) characterize HCP-EHR interaction during care planning in both the prototype and a baseline system.

In this work, we present *ClinicalVis* (Figure 1)[1], an open-source, and freely available visualization-based prototype system with a proof-of-concept design study [58] that we validated with empirical case studies of practicing HCPs and real patient data. ClinicalVis was designed by our team for the purposes of understanding and supporting task-focused interactions between HCPs and EHRs. We compare ClinicalVis to a baseline system, designed to emulate commercially available EHR interfaces that HCPs used in practice at the time of the study. The design study focuses on the use of EHRs for clinical decision making by HCPs for meaningful clinical tasks. Specifically, we use real, anonymized patient EHRs from the MIMIC-III dataset [33] in the two interfaces to observe the HCP-EHR interactions while focused on a realistic clinical scenario — a remote HCP asked to plan care for physiological decompensation amongst multiple patients in an intensive care unit (ICU).

In this paper, we outline related work, present the prototype systems (ClinicalVis and the baseline system), and describe methods used to evaluate HCP interaction with the interfaces. Our findings and insights from are briefly as follows. First, we found that HCP accuracy in forecasting patient needs from EHR was generally poor, and information overload was not overcome by an improved task-agnostic visualization. Second, ClinicalVis improved HCP experience during the tasks, and in a post-task comparison. Finally, we noted that HCPs using ClinicalVis spent their time validating care plans rather than finding information in the EHR, and present considerations for how future work may augment in-situ HCP care planning.

In summary, the contributions of this paper are:

- Providing an open-source EHR visualization-based prototype system for task-focused design evaluation at http://github.com/PAIR-code/clinical-vis.
- Summarizing the findings of a proof-of-concept design study targeting HCP-EHR interactions in a task-focused setting, conducted with real patient data on practicing HCPs.
- Describing insights for future EHR interfaces in presenting clinical data for care planning, and evaluating their efficacy in practice.

[1] http://github.com/PAIR-code/clinical-vis

2 RELATED WORK

We summarize past work on care planning complexity, HCP information needs, evaluation methods for task-focused care planning, and information visualization techniques in EHRs.

Care Planning Complexity in the ICU

Care planning in the ICU is challenging; clinical signals are often irregularly sampled and contaminated by interference and human error. Information visualization and chart reviews become particularly meaningful for decision making in an environment like the ICU where clinicians must process patient information quickly [6], and there is near-constant information from multi-modal devices, and multi-disciplinary staff [63]. Data complexity, and the difficulty of working with EHRs has introduced new threats to patient safety [25], including the introduction of error from fragmented displays and alert fatigue [16]. Prior studies have demonstrated that 80% of "user error" (correlated to 12-22% annual mortality rate) can be attributed to cognitive overload [29, 56]. In this work, we evaluate our prototype system in a realistic task setting to understand the influence of visual EHR summaries in clinical care planning.

Supporting HCP's Information Needs

As the quantity of useful electronic health data burgeons, the ability to efficiently review, interpret, and assimilate diverse data types becomes critical to planning patient care[43]. Current EHR designs contribute to common errors and information loss via mistaken patient identification, mode mismatches, flawed interpretations, incorrect recall or incomplete visibility of system states [60]. Further, EHR usability is poor [32], and contributing factors are not well-understood [20]. EHR software builds in hospitals are often the result of multiple information systems intersecting, with information generated from multiple sources, such as physicians, nurses, billing staff, and even by the patients themselves [27]. Inadequately designed bedside interfaces are the greatest contributor to diagnostic cognitive load [12, 61], and the ability to customize information display for different users and varying clinical scenarios is crucial [47] due to convoluted workflows and prolonged searching activities [48].

Task Based Evaluation of Care Planning

Clinical Decision Support (CDS) systems have been suggested for care planning, but such systems have faced several barriers in adoption. Middleton et al. [44] suggest that CDS dissatisfaction is caused by challenges in aligning an HCP's mental model of the patient, diagnostic process, and therapeutic care plans. Well-designed systems that support decision making steps could bolster CDS integration into

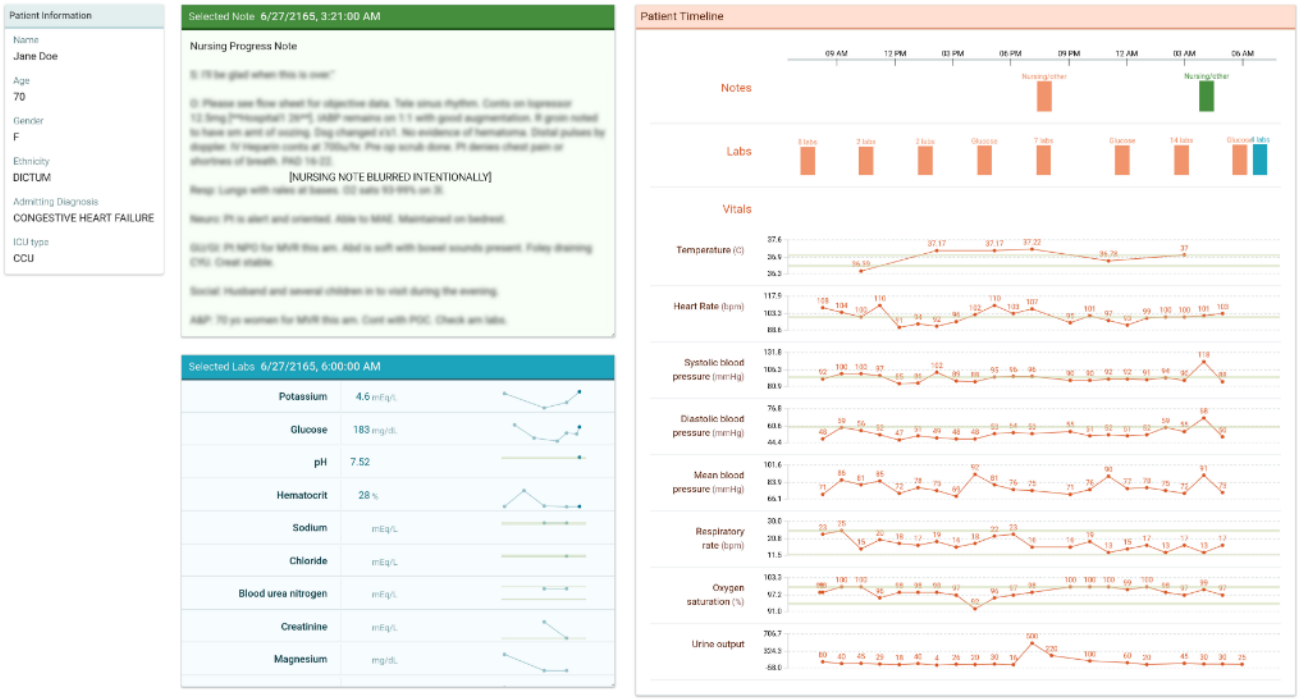

Figure 1: A screenshot of the *ClinicalVis* user interface. Nursing note intentionally obscured to protect privacy. Actual de-identified patient records were displayed during the experiments.

HCP's everyday workflows [54]. Evaluating HCP-EHR interaction is often with Task Load Index (TLX) [26] measures to study the workloads of clinical tasks, and investigate the cognitive demand of EHR transitions [14]. TLX measures have previously been used to study mental, physical and cognitive workload in 17 clinical participants performing a set of clinical tasks on simulated patients in three clinical diagnosis scenarios: urinary tract infection, pneumonia and heart failure [43]. The authors' results suggest that task demands as experienced by HCPs (e.g., needing more clicks, requiring more time) are related to performance (e.g., more severe omission errors) regardless of EHR type. Follow-up work with the same tasks, and quantifying effort towards task and omission errors strengthened these results [45]. In our paper, we specifically target simple prototype designs to focus study on the impact of HCP-EHR interactions during task-based care planning.

Visualization-based Explorations of EHR Design

Information visualization techniques are known to help people carry out tasks more efficiently by highlighting aspects of the data [8] that might require pattern recognition [57], as well as reducing cognitive load and freeing up working memory in decision making [15]. Visual summaries that allow users to explore and analyze their data have a rich and successful standing in infectious disease epidemiology, but these tools are often deeply siloed for specialized applications Carroll et al. [9]. There is also a strong body of prior work on designing visualization-based systems and prototypes that support of teamwork in healthcare coordination [2], and that support patients in managing their own care [4, 37]. There are also many commercial EHR visualizations, such as EPIC and Apple's Health Dashboard. Such work is promising, but duplicating the evaluations conducted on closed-source systems are difficult, and experiments on care planning in a task-oriented setting are hard to reproduce.

There is comparatively little research addressing EHR visualization in a realistic task-focused scenario. Within care planning, implementation of the AWARE EHR visualization system saved time on patient chart review [52] and was associated with improved patient outcomes in the ICU [50], demonstrating that streamlined interfaces can improve the efficiency and reliability of clinical staff [1]. Likewise, systems like LifeLine [53], Timeline [7] and MIVA 2.0 [22], have demonstrated the power of visualizing clinical information visualization using a common time scale for multi-modal clinical data streams. Other systems dynamically scale time

intervals instead so that data scales may be modified by user interaction, e.g., VISITORS [38] and Midgaard [3]. In these cases, systems were evaluated as they are, but not within the focus of a simulated clinical task on real patient data. We differ from these systems in that we focus on simple time-constrained visualizations without any prompts or notifications, to conduct task-focus evaluations of how clinicians move through their workflow through our interface.

3 DESIGN OF PROTOTYPE SYSTEMS

We designed and implemented ClinicalVis (Figure 1), a visualization-based prototype system that supports evaluation of HCP-EHR interaction during realistic clinical care planning tasks using iterative and participatory design methods. Here, we discuss the process and final design of our prototype system.

ClinicalVis Design Process

ClinicalVis is designed as a content-independent but structure-dependent system to enable rapid information assimilation, and support the inference of insights from large amounts of data. We ground our design decisions in known and observed HCP workflows [48], workplace-specific emergent practices [28, 70], and basic requirements outlined in Heath and Luff [28]. We further targeted our designs to mitigate common challenges faced by HCPs in using visual summaries and EHRs [9, 55]. We iteratively stress-tested the system through development, and internally validated our designs against the expertise of our multi-disciplinary team, which includes a practicing physician, machine learning with EHRs expert, and data visualization and interaction design experts.

ClinicalVis Interface and Interactions

ClinicalVis enables HCPs to explore the most recent 24 hours of patient information in four visually-distinct modules, to arrive at a diagnostic care plan for a displayed patient case. The layout and scale of the modules are determined by the physician's reading order, and interaction capabilities are limited to low-level interactions such as scrolling, clicking and hovering. Further, modules can be re-sized for HCP comfort using drag-and-drop.

Clinical notes, ordered labs, and observed vital signs are marked at the time of entry on a 24-hour timeline in a large, single module. Vitals are visualized as a line chart with datapoints at the time of observation, and lab orders and notes are marked as bars at the time of log. The most recent clinical/imaging note and lab is loaded by default in the note- and lab- specific modules, and users can load different notes or labs into appropriate modules by clicking on the 24-hour timeline. The X-axis denotes time moving from left to right, and individual Y-axes denote the max and min for vitals values. Individual labs are visualized as sparklines inside the lab module. Missing or unavailable lab values are not visualized. Although all visualized datapoints for labs and vitals have labels, we included hovering capabilities to enable HCPs to individually read a datapoint's timestamp and numeric value.

Information is color coded by type (vital/lab/note/patient identity) and demographic-agnostic guidelines [31] indicate abnormal values when applicable. Color is sparsely used for encoding at higher levels of abstraction within the modules, and primarily reserved for conveying associations or relaying feedback between different information across modules.

Baseline GUI Design

To understand the influence of ClinicalVis in care planning vis-a-vis commercially available EHRs, we created an imitation control graphical user interface (GUI) as our baseline with the same information and interaction capabilities as ClinicalVis. We model the design of the Baseline GUI (Figure 8) after in-use EHR systems that several authors have first-hand experience with, and knowledge of.

We minimize the influence of any out-of-scope elements by constraining both interfaces to (a) be self-contained on a single page with no external links, (b) support a single patient record at a given time and (c) display a curated set of patient vitals pre-reviewed for task-specific evaluation. During evaluation, HCPs were shown real and unedited clinical data from patient cases in both interfaces. To limit the scope of this study, we exclude any spatial or imaging data (such as X-rays or ultrasound scans), however retaining any associated textual or free-form data.

4 RESEARCH SETTING AND METHODS

Here we describe the study design, data and task choices, participant cohort, and study methods.

We formulated our evaluation to study two complementary research questions:

(1) How do HCPs engage with visual representations of real patient EHR data during task-focused care planning?
(2) Do visual representations of EHRs influence clinical care planning — specifically, do they impact accuracy, confidence and time-to-decision?

Data Source

Clinical data are fundamentally multi-modal, and many different data types are relevant to understanding patient health [66]. We use data from the Medical Information Mart for Intensive Care III (MIMIC-III) database [33]. MIMIC-III is publicly available, and contains over 58,000 hospital admissions from approximately 38,600 adults. Prior visualization work on the open-source MIMIC dataset have focused on web-based tools for researchers, e.g. an interface to identify

cohorts for study creation [39], and predictive mortality modules [11, 40]. Our paper is a novel use of MIMIC-III records to evaluate HPC interaction with EHR systems, and the impact that has on care planning.

Task Definition

We emulated an eICU[2] setting, where multidisciplinary teams of HCPs must forecast care needs to prepare on-site HCPs for therapeutic interventions [10]. We focused on common tasks for physiological decompensation that have potential risks: mechanical ventilation for breathing assistance [65, 68], and vasopressor administration to regulate a patient's blood flow [18, 46].

In our evaluation, we used records of ICU patients aged 15 and older that met two criteria: (a) The patient record did not display any target interventions in the ICU for at least 24 hours before the 8 AM start of "rounds"[3], and for at least 12 hours after rounds, *and* (b) Each patient record had at least 1 note during the 24 hours prior to the rounds.

From this subset, we selected 1 EHR for training, 2 EHRs for the think aloud, and 8 positive and 8 control patients [4] of equal estimated difficulty for the proof-of-concept study as follows:

- **VE+**: First ventilation 4 - 12 hours after rounds.
- **VP+**: First vasopressor 4 - 12 hours after rounds.
- **Control (C)**: No ventilation or vasopressor in the 12 hours after rounds.

Participant Cohort

We recruited 14 clinicians practicing in hospitals in a large metropolitan area who submitted valid responses to a recruitment form. Each potential participant was pre-screened for prior ICU experience (averaging 9 hours per week in the ICU) and the cohort was controlled for diversity of specialization. A summary of our participants' self-reported demographics is available in Table 1. All participants were invited to a laboratory setting for evaluation during the timeframe of the study and successfully completed the study in entirety.

Study Design

We conducted an empirical evaluation of HCP-EHR interaction with ClinicalVis and the baseline prototypes in a mixed-methods, task-focused user study centered around clinical care planning a remote eICU scenario. An overall experimental flow in shown in Figure 2.

Participant sessions lasted approximately 50 minutes, and were conducted in a lab setup at (*redacted for review*). After introducing participants to the study, both prototypes, and the task (5 mins), we conducted a think aloud to qualitatively evaluate the differences in interactions with EHRs between ClinicalVis and the baseline (10 mins). We then conducted a usability study (35 mins) and TLX survey on both interfaces to track how ClinicalVis supported HCPs in care planning. Finally, we used an open-ended comparative survey to capture participant experience (10 minutes).

ID	*Specialization*	*Experience Years*	*ICU Time Hours/Week*
P1	N/A	1+	< 4
P2	Pediatrics	2+	< 4
P3	Infectious diseases	2+	20 to 24
P4	General Surgery	3+	4 to 8
P5	Critical Care	4+	16 to 20
P6	Hospital Medicine	4+	< 4
P7	Emergency Medicine	4+	8 to 12
P8	Internal Medicine	5+	4 to 8
P9	Pediatric Critical Care	5+	16 to 20
P10	Cardiology	5+	16 to 20
P11	Cardiology	5+	< 4
P12	General Medicine	7+	< 4
P13	General Practitioner	10+	< 4
P14	Critical Care	10+	20 to 24

Table 1: Participant breakdown by self-reported specialization, years in current role, and number of hours per week spent in the ICU.

1. Training. At the start of each session, participants were introduced to the goals of the study, the sources of the data, and notified that the data was real de-identified patient information that had not been modified or cleaned. After outlining the three-part structure of the study. participants were introduced to the baseline and ClinicalVis in counterbalanced order on the same patient record. A high-level explanation of the layout and interaction capabilities of both interfaces was provided.

2. Task Introduction. To reduce incentives for providing positively-skewed feedback, we asked participants to perform realistic care-planning in a time-sensitive scenario (possible physical decompensation in each patient record) using the following prompt:

"You are a clinician in an eICU and have just come on-shift. You manage decompensation alerts for two hospitals, each with its own EHR. You are remotely shown 24 hours of the available vitals, labs and notes, and cannot request more. You are asked to review the records of patients from each hospital. Records from Hospital A will look different than those from Hospital B due to the EHR variance. For each

[2]The "eICU" is a clinical term for an "electronic" or remote ICU that a supporting HCP is making judgments from.

[3]Morning rounds are a group evaluation of patient needs and care planning in the ICU.

[4]Demographic details of the selected patients are shown in Table 5.

Figure 2: Experimental flow. Red denotes the Baseline prototype, blue denotes ClinicalVis. Participants were walked through a scripted introduction (1 & 2) to the study, which was divided into three broad sections: A *think-aloud* protocol (3), a task-focused usability study (4) & TLX surveys (5), and a post-completion comparative feedback survey (6). Participants were assigned randomly generated alphanumeric codes as identities to capture data and feedback consistently across the study.

patient, you will decide if staff should be prepared for a Vasopressor, Ventilator, both, or neither in the next 12 hours, and indicate how confident you are."

3. Think aloud. We conducted *think-alouds* to gain qualitative insights into HCP interactions with the baseline and ClinicalVis prototypes, and the usability challenges faced in planning care using both interfaces. [30, 41]. Each participant was shown one pre-selected EHR in counterbalanced order per prototype (also counterbalanced). For each prototype, they were provided with a prompt to think aloud, and a list of assisting questions (see Appendix C). No time limit was enforced, and participants were free to clarify questions about the prototypes, task or patient record.

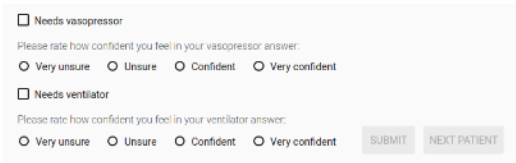

Figure 3: GUI for the evaluation after viewing each patient.

4. Usability Evaluation and 5. TLX Survey. Participating HCPs were assigned to review cases for each prototype sequentially, in counterbalanced order. The ordering of 8 patient records per interface was also counterbalanced using Latin Squares[34]. HCPs were presented with ICU data from the preceding 24 hours of each patient record, and had 120 seconds[5] to arrive at a decision. Resets were permitted in increments of 120 seconds. HCPs were asked to submit a *"yes"* or *"no'* decision for each intervention (vasopressor and ventilator), and indicate their confidence in each decision on a 4-point likert scale with no neutral choice in an evaluation screen 3 at the 2-minute mark. HCPs could open the evaluation screen before the end of the time limit, but could not return to the same case without a reset.

Upon the completion of 8 consecutive patient assessments, participants were asked to complete a survey for the prototype they interacted with. The survey comprised Likert scales measuring mental demand, physical demand, effort level, temporal demand, performance and frustration as defined by the NASA Task Load Index [26] on a 10-point scale.

6. Comparative Survey. Task assessments can lack interpretation, even when observed directly [19]. To capture retrospective feedback after having interacted with both prototypes, participants were asked to compare ClinicalVis and the Baseline along the TLX axes, and optionally provide a reason for their selection. Additionally, participants asked to indicate which of the two prototypes supported the task better. Rationales were solicited as an open-ended free-text responses with no word limit.

Measures. To evaluate these research questions, we collected the following measures:

- **Accuracy (%):** Rate of correct responses for a given case for a participant.
- **Time to Task (seconds):** Time taken to arrive at a decision for a given case. This includes any resets.
- **Confidence:** Self-reported confidence in a decision for a given case using a 4-point Likert scale, scored between -2 (not at all confident) and 2 (very confident).
- **TLX scores:** Self-reported mental demand, physical demand, effort level, hurriedness, success and discouragement on a scale of 0-10.

[5]HCPs often plan care through patient-centered reflective actions [23], and therefore limit the time that each participant has with a patient record.

		By Group			Across All	
		Vasopressor Needed	*Ventilator Needed*	*Control (None)*	*Vasopressor Correct*	*Ventilator Correct*
Accuracy (%)	Baseline	50.00%	56.25%	**71.64%**	62.50%	55.35%
	ClinicalVis	**68.83%**	**62.79%**	67.64%	**63.30%**	**58.92%**
Confidence	Baseline	0.68	0.87	1.34	1.14	0.98
	ClinicalVis	**1.41**	**1.27**	**1.47**	**1.28**	**1.09**
Avg. Time to Task (s)	Baseline	92.31s	92.73s	**83.64s**	87.11s per case/HCP	
	ClinicalVis	**84.43s**	**86.86s**	85.37s	**85.94s** per case/HCP	

Table 2: Accuracy, confidence and time-to-task results of average participant performance per case using the Baseline and ClinicalVis prototypes. The best results for each comparison are bolded.

5 RESULTS AND DISCUSSION

In this section we discuss our findings and elaborate on how those suggest design implications for future work on EHRs and HCP-EHR interaction, discussing these results in turn. Table 2 summarizes our quantitative results from 224 simulated EHR-HCP encounters (112 for each prototype).

Our key insights are primarily related to practice, specifically that (a) EHR to task accuracy in care planning is generally poor, and, counter-intuitively, did not improve with a better visualization; (b) visualizations changed the way HCPs experience data; and (c) HCPs maximized time for validation when interacting with visual summaries.

EHR-To-Task Performance is Generally Poor

Our hypothesis was that an improved visual interface would lead to better HCP performance on planning care for realistic tasks, where better performance was characterized by increased accuracy and/or lowered time-to-decision. Our analysis revealed that overall performance was better on ClinicalVis than on the Baseline, but insignificantly so.

Participant's average accuracy was higher when using ClinicalVis (63.30%) compared to the baseline (62.50%), and overall accuracy remained under 75%. Further, there was no observable pattern in accuracy across individual participant performance for the two interfaces (Fig4).

Though self-reported HCP confidence in planned care was significantly higher when using ClinicalVis (1.18 vs 1.06 for baseline), the average time-to-decision was insignificantly lower in ClinicalVis (85.94s/case/HCP) than the baseline (87.11/case/HCP).

EHR Are Fundamentally Limited For Care Planning. While prior work has found clinical and economic benefits in adopting eICUs [17], many of the signals factored into care planning come from being able to physically interact with the patient [10].

"*Part of the frustration with EHRs in general, is that people try and make predictions based on the data, which is just not as helpful as laying eyes on the patient*" - P4

Seven participants echoed that not seeing the patient reduced their confidence considerably and this influenced the amount of caution exercised, with one participant estimating that "*90% of signals come from physically observing the patient*". Further, in the absence of the patient and/or data, seven participants reported that they were either re-reading, over-thinking or reading too much into the admitting diagnosis, EHR or prognosis.

"*eICUs are actually of questionable effectiveness, because there is really only so much info that you can get from the data. With any one of these examples, my confidence would go up considerably if I actually saw the patient.*" - P4

While our work addresses the need to integrate well within an HCP's workflow, our findings show that even when evaluating individual cases, an HCP's information needs are complex and constantly changing. The physical absence of

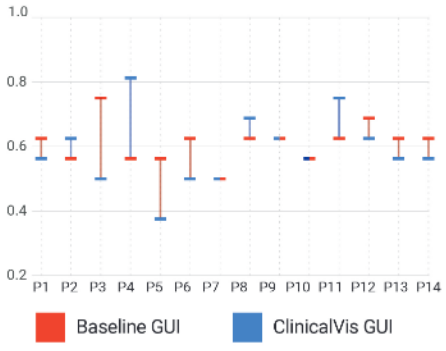

Figure 4: Individual HCP accuracy in care planned with Baseline and ClinicalVis, higher is better.

the patient, unavailability of data and inconsistencies within the data were cited as key causes for frustrations and delays in interacting with EHRs on both interfaces.

Fundamental Data Concerns that affect HCPs. Participant feedback indicated that the key factors influencing clinical preparedness in our evaluation were rooted in the underlying EHR data. Many EHRs have *incomplete* or *inconsistent* data in the underlying source record, reflecting actual available information [24]. Sources for these include different intervals for aggregation, discrepancies between the time of an observation and time of logging the observation into the system, and hospital- or ICU- specific practices.

"(It is) The nature of data that prevents me from making a decision, such as lack of knowledge of interventions." - P5

"I look at what's going on in the last 4 hours because nurses won't put everything at the same time though they try." - P1

HCPs interacted with prototype systems with reduced confidence in cases *they perceived* as having insufficient data. Inconsistencies in data presentation and non-standardized data logging practices further exacerbated frustrations and reduced confidence during interactions with an EHR for both interfaces. For instance, five participants requested that urine output be as an aggregate rather than in absolute units - however, each indicated different preferences for the time interval (1 hour, 12 hours, 24 hours) for aggregation. Additionally, two participants informed us that they were unfamiliar with abbreviations in the notes. One participant observed that *"notes written in all uppercase make me believe that the nurse is yelling at me."*

Better Visualization Improved HCP Experience

Participants reported that ClinicalVis had little impact on their overall performance, but they experienced the data differently. In line with our design goals, we found that participants preferred ClinicalVis over the Baseline. They reported lower cognitive loads, reduced opportunities for error, increased content engagement and information retrieval.

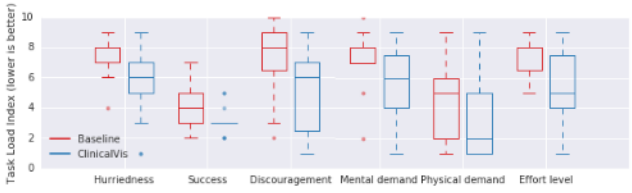

Figure 5: Individual responses to each prototype for the TLX survey. Lower scores are better, and ClinicalVis (blue) consistently had improved responses.

TLX analysis. A stronger preference for ClinicalVis was evident on all TLX dimensions (Figure 5). We computed the the Kolmogorov-Smirnov statistic [42] on the two samples (ClinicalVis and baseline) to test for significance in each question. We found that ClinicalVis passed statistical significance ($p = 0.10$) level for feelings of discouragement ($p = 0.017$) and effort ($p = 0.051$). However, at this level, participant preference for ClinicalVis on mental demand, physical demand, hurriedness, and success was not statistically significant.

Prior work has found that EHR-related task demands significantly increases mental effort and task difficulty, which are predictors of omission error performance [45]. One participant remarked that *"[I] felt reluctant reading the table [baseline] and going into the minor details... I was more comfortable making decisions having engaged with the data at a greater level [in the visualization]."* (P13). Increased loads reported by the HCPs for the baseline can be viewed as a proxy for the number of general omission errors likely to be made by systems in practice.

Experiencing EHRs differently. We observed that ClinicalVis integrated well into physician work flow, and physicians reported being able to establish clearer clinical narratives, find responses to questions faster, and identify outliers and relevance with lower effort. Participants were able to quickly habituate themselves to our prototype system. Interactions were learned quickly, and all 14 participants reported spending less attention on the interface and more on the data. In contrast, nine participants described their performance on the last few patient records in the baseline as poor due to "mental fatigue", "excessive effort" or "exhaustion". One participant stated that *"the baseline felt very demanding for patients that were not very sick"*, while also noting that *"This one (CinicalVis) felt slower, I came to a decision faster so I tried to slow myself down and I think I was over-thinking it, but it was very clear if the patient was going to need a vasopressor or a ventilator or if they were going to decompensate."*

Data-first interface. Given the distributed nature of tasks in an ICU, HCPs frequently rely on interpersonal trust within and between the ICU team and consulting services to plan care [69]. In the absence of such support, HCPs felt varying degrees of frustration, effort and assistance with both prototypes (Fig 6). HCPs preferred ClinicalVis to the baseline across four different metrics: 1) which visualization made the task feel more rushed, 2) which required more work in completing the task, 3) which was most frustrating to use, and 4) which was a better at supporting the task. Two participants felt no difference in effort across the two interfaces, instead attributing the effort applied to the patient cases.

Altered perception of content . During the comparative evaluation, participants tended to focus on the increased mental demand required from the baseline, noting that they felt rushed and anxious about finding relevant information:

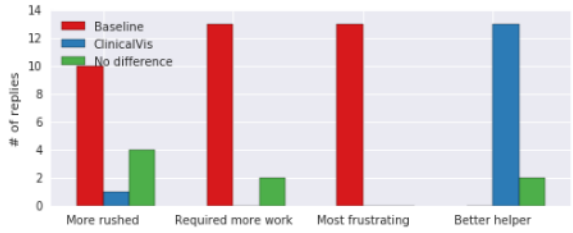

Figure 6: Post-study comparative evaluation of the Baseline and ClinicalVis on TLX dimensions. Note that in the first three comparisons of *negative* affect the Baseline scores higher. In the final comparison of *positive* affect, ClinicalVis scores highest.

"Certain sense of anxiety... trying to make these predictions is hard, and when you feel like you're having to fight with the UI to find what you are looking for.'. When asked about factors determining the indicated confidence levels in the baseline, the general sentiment pointed towards a fear of having missed a number, or a lack of confidence in the data itself. Such concerns are well-established in prior work on HCP-EHR interaction, in which a majority of surveyed clinicians were worried about overlooking important information due to the volume of data and inadequate display/organization [48]. In contrast, the ClinicalVis system allowed them to have more confidence in their ultimate assessments of the patient. *"Some interfaces in hospital are (pause) unfortunately like the baseline and they make you feel like you have no control. It makes you nervous. The visualization is much more reassuring."*

Participants' comments about the baseline were primarily centered around finding and perceiving information, whereas feedback on ClinicalVis tended to describe the status of the patient, indicating deeper engagement with the data. Participants had a tendency to use terms such as "*spikes*", "*dips*", "*trending up*" to describe the patient data rather than describing the visualization, suggesting that information was being consumed at a faster rate. Further, all participants claimed that notes was easier to read and four participants assumed that it had been extrinsically modified to improve reading, though they had not.

Clinical Confidence in Visualization. Prior work has indicated that HCPs primarily focused on the viewing of clinical notes during electronic chart reviews (ECR) in the ICU (44.4% of their time), followed by laboratories (13.3%), imaging studies (11.7%), and searching/scrolling (9.4%) for typical cases [48]. 47% of ECRs also began with review of clinical notes, which were the most common navigation destination.

Participant's interactions with the data in our interface echoed these findings, with two exceptions. First, all physicians started with reviewing patient information. We can speculate that additional attention was paid to this module due to the influence of the study and the absence of the physical patient. Secondly, the review of clinical notes was heavily interspersed with the viewing of vitals and labs. In particular, physicians looked to clinical notes to iteratively answer questions that arose from the charts, and moved quickly between the patient timeline module and the notes module. Two participants were observed contextualizing parts of the nursing note within the patient timeline, explaining that notes were typically created over the course of a few hours and logged at a later time. In such scenarios, we find that a single-screen and modular approach to information presentation creates tighter feedback cycles, enabling HCPs to confirmed their intuitions with comments in the notes. *"The graphic interface is much more helpful, the separation of subjective/ objective, and the trending function is just better in the visualization."* When asked about effort, participants emphasized the difficulty of locating information in the baseline. *"It's harder to find the information that you are looking for, harder to see trends and separate out the different components that you are looking at."*

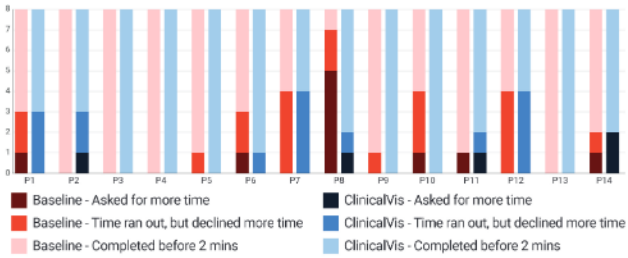

Figure 7: Time requirements per participant for all tasks in the Baseline compared to ClinicalVis. Note that HCPs often asked for more time (dark red) when using the Baseline.

Maximizing Time For Decision Validation, Not Information Finding

In time-sensitive environments, the efficient use of time is paramount to care planning. Maximizing the efficiency of time and attention spent with each patient record in ClinicalVis allowed participants to verify facts contributing to care plans.

Inadequacies of Time to Task metric. HCPs spent time in *validating their decisions* in ClinicalVis, compared to time spent in *looking for information* in the baseline (Fig 7). *"Because I spent more time looking... so I had less time with the table.'*. We found that ClinicalVis gave participants a sense that they were able to digest data more completely, and time remaining after having arrived at a decision was used for validation as opposed to submitting in the evaluation screen. In the light of this behavior, the insignificant difference between the average time to task per case per HCP between the

two interfaces suggests that time to task was an inadequate measure of performance.

Influence of individual assessments of caution. We anticipated that HCPs care about treating the patient, through patient-centered reflective actions [23], and therefore may increase cautious planning. However, prior work has been unable to prove that routine reflection in action can improve diagnostic outcomes[49]. In our study, we find that while providing avenues in EHR interfaces that support active reading and routine reflection did not increase outcome accuracy, the increased time-to-task and reduced accuracy for control cases suggests a higher false positive rate. Participant descriptions attribute this to increased levels of caution and clinical preparedness for "a prognosis that could go either way".

It was observed that participating physicians exercised varying levels of caution in planning care when unsure, with some being more conservative than others. Some physicians were more cautious and willing to prepare resources regardless of confidence in their decision. For instance, in one case, a participant noted that they felt unsure of prognosis, even though the patient appeared stable — the participant chose to prepare for vasopressor administration. For the same case, another participant was "*confident how this (prognosis) would go*" and did not prepare for either, correctly.

Validation practices that support heuristic methods. Prior work has indicated that HCPs apply a hypothetico-deductive method for arriving at a diagnosis during a patient encounter [21], in which diagnostic hypotheses are "proposed, tested, and either verified or rejected". This is further augmented by various "mental shortcuts" or heuristics that can sometimes lead to diagnostic errors. [54]. We observed several of these heuristics in practice when using ClinicalVis and the baseline. Particularly, HCPs frequently exhibited the use of the representative heuristic[6] in both interfaces. However, the related heuristics of anchoring [7], premature closure[8] and confirmation bias[9] appeared to be mitigated through data validation practices.

HCPs were frequently observed to verbally change their decisions in ClinicalVis as they traversed different parts of the visualization; further, they demonstrated better recall of patient information in the evaluation screen; finally, twelve participants noted that it was easier to establish clinical narrative using ClinicalVis. The validation experience is a critical because reflection and communication are parts of clinical experience that technology currently hinders, and prior work has demonstrated that 91% of all medical errors can be attributed to difficulties in communication and collaboration [13].

[6]Used to estimate the likelihood of a condition based on how closely a patient's presentation matches prototypical cases of the condition. [54]
[7]A tendency to be attached to initial diagnostic hypothesis despite the accumulation of contradictory evidence.
[8]Settling on a diagnosis with insufficient investigation of information.
[9]Seeking evidence that supports working hypotheses, ignoring contradictory or ambiguous evidence.

6 LIMITATIONS AND FUTURE WORK

Our work is a first step towards an evaluation of visualization systems that focuses on the task-oriented setting in which clinicians work, and the environmental factors they face. Here we detail limitations and the future work that is needed.

First, our work limited the displayed data to the most direct indicators for the chosen care planning tasks. A more robust exploration of data sources and types should be integrated into the ClinicalVis system, accounting for timescale and sampling differences. Second, we created ClinicalVis with simple design principles, without an extensive design-focused process. Future work is needed to investigate whether visual representations that mimic denotations and reading practices of HCPs would improve the evaluated measures and outcomes. Third, we investigated two specific tasks in a specific care planning setting. More complete sets of decision making situations should be tested — both in terms of the tasks and scenarios. Fourth, our sample of HCPs was limited, and we believe that a larger sample size of HCPs should be examined to validate our findings for a broader population. Finally, our finding that accuracy did not improve with better visualization of patient data was interesting and unexpected, and should be studied further.

7 CONCLUSION

In this paper, we present an empirical evaluation of ClinicalVis, a visualization-based prototype system, on the interactions of HCPs with EHRs in a task-focused setting. We then present insights learned from 14 participating HCPs as they interacted with EHRs in a simulated eICU to plan for care for real patient cases using ClinicalVis. Specifically, we found that (a) counter-intuitively, EHR-to-task was generally poor in the physical absence of patients, (b) ClinicalVis positively altered the way HCPs experienced data without significantly impacting performance ,and (c) physicians preferred to maximize available time by purposing it for decision-validation in ClinicalVis. Importantly, HCPs using ClinicalVis reported increased confidence, which is particularly relevant as the field of machine learning begins to target the use of EHR data to answer specific clinical questions [24, 62, 67].

While we chose data that was most appropriate for predicting physiological decompensation, our proof-of-concept interface is agnostic to the specific experimental task performed. ClinicalVis is scalable to include variables other than the ones presented; our visualization is not custom-fit for this task, and can be used more generally to investigate during

and post-task clinical usage of ICU EHR. The design implications of our work suggest that modeling clinical data for decision support should include elements to guide clinical use, and to that end, we have open-sourced ClinicalVis as a means to support and engender such efforts. We believe that ClinicalVis can open up avenues for the rigorous evaluation of interactions between clinicians and patient data to best improve aspects of healthcare delivery.

ACKNOWLEDGMENTS

We would like to thank several people who contributed to making this work substantially better, including Lauren Dukes, Michael Terry, and Joseph Jay Williams.

REFERENCES

[1] Adil Ahmed, Subhash Chandra, Vitaly Herasevich, Ognjen Gajic, and Brian W Pickering. 2011. The effect of two different electronic health record user interfaces on intensive care provider task load, errors of cognition, and performance. *Critical care medicine* 39, 7 (2011), 1626–1634.

[2] Ofra Amir, Barbara J Grosz, Krzysztof Z Gajos, Sonja M Swenson, and Lee M Sanders. 2015. From care plans to care coordination: Opportunities for computer support of teamwork in complex healthcare. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1419–1428.

[3] Ragnar Bade, Stefan Schlechtweg, and Silvia Miksch. 2004. Connecting time-oriented data and information to a coherent interactive visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 105–112.

[4] Stinne Aaløkke Ballegaard, Thomas Riisgaard Hansen, and Morten Kyng. 2008. Healthcare in everyday life: designing healthcare services for daily life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1807–1816.

[5] David W Bates. 2010. Getting in step: electronic health records and their role in care coordination.

[6] James M Blum and Kevin K Tremper. 2010. Alarms in the intensive care unit: too much of a good thing is dangerous: is it time to add some intelligence to alarms? *Critical care medicine* 38, 2 (2010), 702–703.

[7] Alex AT Bui, Denise R Aberle, and Hooshang Kangarloo. 2007. TimeLine: visualizing integrated patient records. *IEEE Transactions on Information Technology in Biomedicine* 11, 4 (2007), 462–473.

[8] Stuart K Card, Jock D Mackinlay, and Ben Shneiderman. 1999. *Readings in information visualization: using vision to think*. Morgan Kaufmann.

[9] Lauren N Carroll, Alan P Au, Landon Todd Detwiler, Tsung-chieh Fu, Ian S Painter, and Neil F Abernethy. 2014. Visualization and analytics tools for infectious disease epidemiology: a systematic review. *Journal of biomedical informatics* 51 (2014), 287–298.

[10] Leo Anthony Celi, Erkan Hassan, Cynthia Marquardt, Michael Breslow, and Brian Rosenfeld. 2001. The eICU: itâĂŹs not just telemedicine. *Critical care medicine* 29, 8 (2001), N183–N189.

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

[12] David Chou. 2012. Health IT and patient safety: building safer systems for better care. *Jama* 308, 21 (2012), 2282–2282.

[13] Trevor Cohen, Brett Blatter, Carlos Almeida, Edward Shortliffe, and Vimla Patel. 2006. A cognitive blueprint of collaboration in context: Distributed cognition in the psychiatric emergency department. *Artificial intelligence in medicine* 37, 2 (2006), 73–83.

[14] Lacey Colligan, Henry WW Potts, Chelsea T Finn, and Robert A Sinkin. 2015. Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. *International journal of medical informatics* 84, 7 (2015), 469–476.

[15] Kristin A Cook and James J Thomas. 2005. Illuminating the path: The research and development agenda for visual analytics. (2005).

[16] National Research Council et al. 2012. Health IT and patient safety: building safer systems for better care.

[17] Alberto Coustasse, Stacie Deslich, Deanna Bailey, Alesia Hairston, and David Paul. 2014. A business case for tele-intensive care units. *The Permanente Journal* 18, 4 (2014), 76.

[18] Frederick D'Aragon, Emilie P Belley-Cote, Maureen O Meade, François Lauzier, Neill KJ Adhikari, Matthias Briel, Manoj Lalu, Salmaan Kanji, Pierre Asfar, Alexis F Turgeon, et al. 2015. Blood Pressure Targets For Vasopressor Therapy: A Systematic Review. *Shock* 43, 6 (2015), 530–539.

[19] Alan Dix. 2009. Human-computer interaction. In *Encyclopedia of database systems*. Springer, 1327–1331.

[20] Marc A Ellsworth, Mikhail Dziadzko, John C OâĂŹHoro, Ann M Farrell, Jiajie Zhang, and Vitaly Herasevich. 2016. An appraisal of published usability evaluations of electronic health records via systematic review. *Journal of the American Medical Informatics Association* 24, 1 (2016), 218–226.

[21] Arthur S Elstein, Lee S Shulman, and Sarah A Sprafka. 1978. Medical problem solving an analysis of clinical reasoning. (1978).

[22] P Faiola A, Srinivas and J Duke. 2015. Supporting Clinical Cognition: A Human-Centered Approach to a Novel ICU Information Visualization Dashboard. In *AMIA Annu Symp Proc*. AMIA, 560–569.

[23] Gerhard Fischer. 2001. Communities of interest: Learning through the interaction of multiple knowledge systems. In *Proceedings of the 24th IRIS Conference*, Vol. 1. Department of Information Science, Bergen, 1–13.

[24] Marzyeh Ghassemi, Tristan Naumann, Peter Schulam, Andrew L Beam, and Rajesh Ranganath. 2018. Opportunities in Machine Learning for Healthcare. *arXiv preprint arXiv:1806.00388* (2018).

[25] Yong Y Han, Joseph A Carcillo, Shekhar T Venkataraman, Robert SB Clark, R Scott Watson, Trung C Nguyen, Hülya Bayir, and Richard A Orr. 2005. Unexpected increased mortality after implementation of a commercially sold computerized physician order entry system. *Pediatrics* 116, 6 (2005), 1506–1512.

[26] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[27] Kristiina Häyrinen, Kaija Saranto, and Pirkko Nykänen. 2008. Definition, structure, content, use and impacts of electronic health records: A review of the research literature. *International journal of medical informatics* 77 5 (2008), 291–304.

[28] Christian Heath and Paul Luff. 1996. Documents and professional practice:âĂIJbadâĂİ organisational reasons for âĂIJgoodâĂİ clinical records. In *Proceedings of the 1996 ACM conference on Computer supported cooperative work*. ACM, 354–363.

[29] Kerm Henriksen, Elizabeth Dayton, Margaret A Keyes, Pascale Carayon, and Ronda Hughes. 2008. Understanding adverse events: a human factors framework. (2008).

[30] Andreas Holzinger. 2005. Usability engineering methods for software developers. *Commun. ACM* 48, 1 (2005), 71–74.

[31] Gary L Horowitz, Sousan Altaie, and James C Boyd. 2010. *Defining, establishing, and verifying reference intervals in the clinical laboratory; approved guideline*. CLSI.

[32] Howe JL, Adams KT, Hettinger A, and Ratwani RM. 2018. Electronic health record usability issues and potential contribution to patient harm. *JAMA* 319, 12 (2018), 1276–1278. https://doi.org/10.1001/jama.2018.1171

[33] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3 (2016).

[34] A Donald Keedwell and József Dénes. 2015. *Latin squares and their applications*. Elsevier.

[35] Kathy Kenyon. [n. d.]. Overcoming Contractual Barriers To EHR Research. https://www.healthaffairs.org/do/10.1377/hblog20151014.051141/full/

[36] Saif Khairat, George Cameron Coleman, Samantha Russomagno, and David Gotz. [n. d.]. Assessing the Status Quo of EHR Accessibility, Usability, and Knowledge Dissemination. ([n. d.]).

[37] Predrag Klasnja, Andrea Civan Hartzler, Kent T Unruh, and Wanda Pratt. 2010. Blowing in the wind: unanchored patient information work during cancer care. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 193–202.

[38] Denis Klimov, Yuval Shahar, and Meirav Taieb-Maimon. 2010. Intelligent visualization and exploration of time-oriented data of multiple patients. *Artificial intelligence in medicine* 49, 1 (2010), 11–31.

[39] Joon Lee, Evan Ribey, and James R Wallace. 2015. A web-based data visualization tool for the MIMIC-II database. *BMC medical informatics and decision making* 16, 1 (2015), 15.

[40] Dina Levy-Lambert, Jen J Gong, Tristan Naumann, Tom J Pollard, and John V Guttag. 2018. Visualizing Patient Timelines in the Intensive Care Unit. *arXiv preprint arXiv:1806.00397* (2018).

[41] Clayton Lewis and John Rieman. 1993. Task-centered user interface design. *A Practical Introduction* (1993).

[42] Frank J Massey Jr. 1951. The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association* 46, 253 (1951), 68–78.

[43] Lukasz M Mazur, Prithima R Mosaly, Carlton Moore, Elizabeth Comitz, Fei Yu, Aaron D Falchook, Michael J Eblan, Lesley M Hoyle, Gregg Tracton, Bhishamjit S Chera, et al. 2016. Toward a better understanding of task demands, workload, and performance during physician-computer interactions. *Journal of the American Medical Informatics Association* 23, 6 (2016), 1113–1120.

[44] B Middleton, DF Sittig, and A Wright. 2016. Clinical Decision Support: a 25 Year Retrospective and a 25 Year Vision. *Yearbook of medical informatics* (2016), S103.

[45] Prithima Reddy Mosaly, Lukasz M Mazur, Fei Yu, Hua Guo, Merck Derek, David H Laidlaw, Carlton Moore, Lawrence B Marks, and Javed Mostafa. 2018. Relating task demand, mental effort and task difficulty with physicians' performance during interactions with electronic health records (EHRs). *International Journal of Human–Computer Interaction* 34, 5 (2018), 467–475.

[46] Marcus Müllner, Bernhard Urbanek, Christof Havel, Heidrun Losert, Gunnar Gamper, and Harald Herkner. 2004. Vasopressors for shock. *The Cochrane Library* (2004).

[47] Matthew E Nolan, Rodrigo Cartin-Ceba, Pablo Moreno-Franco, Brian Pickering, and Vitaly Herasevich. 2017. A Multisite Survey Study of EMR Review Habits, Information Needs, and Display Preferences among Medical ICU Clinicians Evaluating New Patients. *Applied clinical informatics* 8, 04 (2017), 1197–1207.

[48] Matthew E Nolan, Rizwan Siwani, Haytham Helmi, Brian W Pickering, Pablo Moreno-Franco, and Vitaly Herasevich. 2017. Health IT Usability Focus Section: Data Use and Navigation Patterns among Medical ICU Clinicians during Electronic Chart Review. *Applied clinical informatics* 8, 04 (2017), 1117–1126.

[49] Geoff Norman. 2009. Dual processing and diagnostic errors. *Advances in Health Sciences Education* 14, 1 (2009), 37–49.

[50] Natalia Olchanski, Mikhail A Dziadzko, C Tiong, Craig E Daniels, Steve G Peters, John C O'Horo, and Michelle N Gong. 2017. Can a Novel ICU Data Display Positively Affect Patient Outcomes and Save Lives? *Journal of medical systems* 41, 11 (2017), 171.

[51] Ann S OâĂŹmalley, Joy M Grossman, Genna R Cohen, Nicole M Kemper, and Hoangmai H Pham. 2010. Are electronic medical records helpful for care coordination? Experiences of physician practices. *Journal of general internal medicine* 25, 3 (2010), 177–185.

[52] Brian W Pickering, Yue Dong, Adil Ahmed, Jyothsna Giri, Oguz Kilickaya, Ashish Gupta, Ognjen Gajic, and Vitaly Herasevich. 2015. The implementation of clinician designed, human-centered electronic medical record viewer in the intensive care unit: a pilot step-wedge cluster randomized trial. *International journal of medical informatics* 84, 5 (2015), 299–307.

[53] Catherine Plaisant, Brett Milash, Anne Rose, Seth Widoff, and Ben Shneiderman. 1996. LifeLines: visualizing personal histories. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. ACM, 221–227.

[54] Alvin Rajkomar and Gurpreet Dhaliwal. 2011. Improving diagnostic reasoning to improve patient safety. *The Permanente Journal* 15, 3 (2011), 68.

[55] Lisette C Roman, Jessica S Ancker, Stephen B Johnson, and Yalini Senathirajah. 2017. Navigation in the electronic health record: a review of the safety and usability literature. *Journal of biomedical informatics* 67 (2017), 69–79.

[56] Jeffrey M Rothschild, Christopher P Landrigan, John W Cronin, Rainu Kaushal, Steven W Lockley, Elisabeth Burdick, Peter H Stone, Craig M Lilly, Joel T Katz, Charles A Czeisler, et al. 2005. The Critical Care Safety Study: The incidence and nature of adverse events and serious medical errors in intensive care. *Critical care medicine* 33, 8 (2005), 1694–1700.

[57] Anya Savikhin, Ross Maciejewski, and David S Ebert. 2008. Applied visual analytics for economic decision-making. In *Visual Analytics Science and Technology, 2008. Vast'08. Ieee Symposium on*. IEEE, 107–114.

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

[59] Angela Shen-Hsieh and Mark Schindl. 2002. Data visualization for strategic decision making. In *Case Studies of the CHI2002*. ACM, 1–17.

[60] Ben Shneiderman, Catherine Plaisant, and Bradford W Hesse. 2013. Improving healthcare with interactive visualization. *Computer* 46, 5 (2013), 58–66.

[61] C Sinsky, J Hess, BT Karsh, JP Keller, and R Koppel. 2012. Comparative user experiences of health IT products: how user experiences would be reported and used. *Institute of Medicine of the National Academies* (2012).

[62] Andrew J Steele, S Aylin Cakiroglu, Anoop D Shah, Spiros C Denaxas, Harry Hemingway, and Nicholas M Luscombe. 2018. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *bioRxiv* (2018), 256008.

[63] Kathleen M Sutcliffe, Elizabeth Lewton, and Marilynn M Rosenthal. 2004. Communication failures: an insidious contributor to medical mishaps. *Academic Medicine* 79, 2 (2004), 186–194.

[64] Darius Tahir. [n. d.]. Doctors barred from discussing safety glitches in U.S.-funded software. https://www.politico.com/story/2015/09/doctors-barred-from-discussing-safety-glitches-in-us-funded-software-213553

[65] Martin J Tobin (Ed.). 2006. *Principles and practice of mechanical ventilation.* McGraw-Hill Medical Pub. Division.

[66] Griffin M Weber, Kenneth D Mandl, and Isaac S Kohane. 2014. Finding the missing link for big biomedical data. *Jama* 311, 24 (2014), 2479–2480.

[67] Jionglin Wu, Jason Roy, and Walter F Stewart. 2010. Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches. *Medical care* (2010), S106–S113.

[68] Karl L Yang and Martin J Tobin. 1991. A prospective study of indexes predicting the outcome of trials of weaning from mechanical ventilation. *New England Journal of Medicine* 324, 21 (1991), 1445–1450.

[69] Jiajie Zhang, Vimla L Patel, Kathy A Johnson, and Jack W Smith. 2002. Designing human-centered distributed information systems. *IEEE intelligent systems* 17, 5 (2002), 42–47.

[70] Xiaomu Zhou, Mark S Ackerman, and Kai Zheng. 2009. I just don't know why it's gone: maintaining informal information use in inpatient care. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 2061–2070.

A BASELINE OVERVIEW

The baseline visualization is a composite representation of a range of real commercial EHRs that the research team have directly observed.

B DATASET STATISTICS

From a clinical perspective, the data we present in these interfaces can be described as a mix of categorical variables, numerical variables and textual or free-form data. Nominal categorical variables include demographic information about the patient, admitting diagnosis and ICU type. Additional ordered and unordered categorical data are also contained within free-form nursing and radiology notes. Numerical data are often seen in labs and vitals, in which they are presented either on an interval scale or a ratio scale. All observed variables have temporal attributes associated with them, either per observation (vitals and labs) or for a set of observations (as is in the case of nursing and radiology notes).

C STUDY PROMPTS

Thinkaloud Prompt:

> Please walk me through what you see on the screen, and verbalize any thoughts that you have as you arrive at a decision for this patient.

Assistive Questions:

- What you believe is happening here?
- Are you looking for something specific?
- What are you looking for?
- What action are you trying to perform?
- Why you are trying to perform this action?
- What do you expect will happen?

Static Variables	Gender
	Age
	Ethnicity
	ICU
	Admission Type
Vitals and Labs	Anion gap
	Bicarbonate
	Blood pH
	Blood urea nitrogen
	Chloride
	Creatinine
	Diastolic blood pressure
	Fraction inspired oxygen
	Glascow coma scale total
	Glucose
	Heart rate
	Hematocrit
	Hemoglobin
	INR*
	Lactate
	Magnesium
	Mean blood pressure
	Oxygen saturation
	Partial thromboplastin time
	Phosphate
	Platelets
	Potassium
	Prothrombin time
	Respiratory rate
	Sodium
	Systolic blood pressure
	Temperature
	Weight
	White blood cell count
	Phosphorus

*International normalized ratio of the prothrombin time

Table 3: Variables included in ClinicalVis.

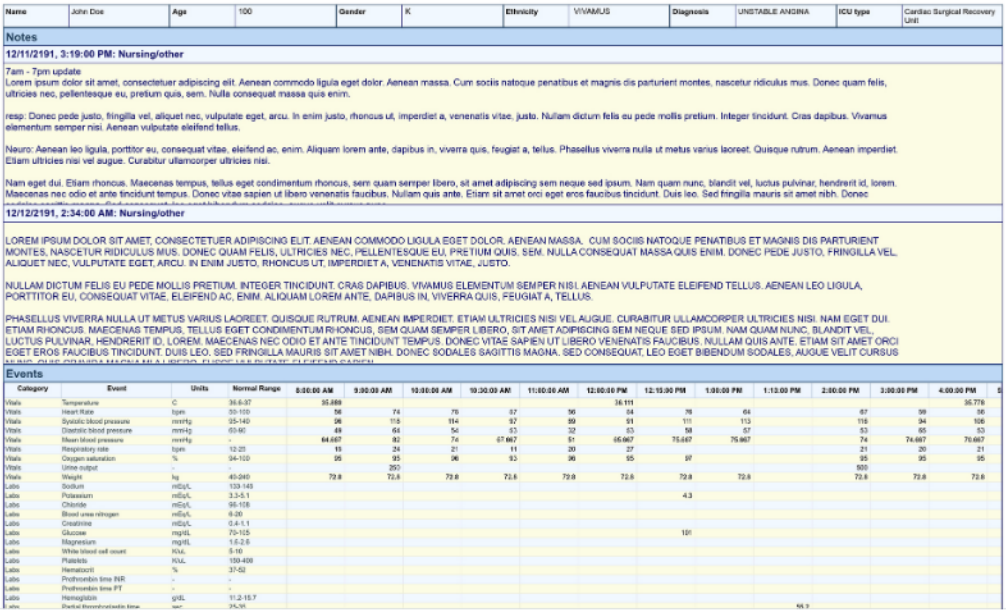

Figure 8: The baseline user interface displaying fake patient data. Note that real patient data was shown to all participants during the study, but cannot be shown here.

Labels	ID	Gender	Age	ICU
Control	13212	M	59	CCU
	14474	M	74	MICU
	14593	M	77	CCU
	5268	F	56	SICU
	59381	F	45	SICU
	69857	M	84	MICU
	9130	F	60	MICU
	91513	F	57	SICU
VP+	32099	F	37	TSICU
	7479	M	61	MICU
VE+	28940	F	61	MICU
	48038	F	33	SICU
VP+ and VE+	1115	M	73	CSRU
	14495	M	54	CCU
	21454	F	70	CCU
	5285	F	54	CSRU
Think Aloud	25328	F	78	CCU

Table 4: Selected Patients for the task. ID refers to the subject's MIMIC-III subject identifier. The gender, age and admitting ICU are also reported. Abbreviations: MICU, medical care unit; SICU, surgical care unit; TSICU, trauma surgical care unit; CCU, cardiac care unit; CSRU, cardiac-surgery recovery unit.

	Control	VP+	VE+	Both
Female	76	1	9	23
Male	137	2	12	52
CCU	76	1	5	38
CSRU	29	0	4	20
MICU	45	1	6	5
NICU	1	0	0	0
SICU	41	0	4	5
TSICU	21	1	2	7
Average Age	72	47	63	68

Table 5: Demographics of all eligible patients for the task. VP+ refers to patients who received vasopressors. VE+ refers to patients who received invasive ventilation. Abbreviations: MICU, medical care unit; SICU, surgical care unit; TSICU, trauma surgical care unit; CCU, cardiac care unit; CSRU, cardiac-surgery recovery unit.