Journal of Biomedical Informatics 132 (2022) 104107

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Tell me something interesting: Clinical utility of machine learning prediction models in the ICU

Bar Eini-Porat [a,*], Ofra Amir [a], Danny Eytan [a,b], Uri Shalit [a]

[a] Technion - Israel Institute of Technology, Haifa, Israel
[b] Critical Care Unit, Rambam Medical Center, Technion - Israel Institute of Technology, Haifa, Israel

ARTICLE INFO

Keywords
ICU
Machine learning
Decision-support
Vital signs
2022 MSC
00-01
99-00

ABSTRACT

In recent years, extensive resources are dedicated to the development of machine learning (ML) based clinical prediction models for intensive care unit (ICU) patients. These models are transforming patient care into a collaborative human-AI task, yet prediction of patient-related events is mostly treated as a standalone goal, without considering clinicians' roles, tasks or workflow in depth. We conducted a mixed methods study aimed at understanding clinicians' needs and expectations from such systems, informing the design of machine learning based prediction models. Our findings identify several areas of focus where clinicians' needs deviate from current practice, including desired prediction targets, timescales stemming from actionability requirements, and concerns regarding the evaluation and trust in these algorithms. Based on our findings, we suggest several design implications for ML-based prediction tools in the ICU.

1. Introduction

The intensive care unit (ICU) provides specialized care for critically ill patients with life-threatening conditions. Decisions are made constantly based on an ongoing flow of physiological readings and ancillary patient data. The abundance of information and rising workloads are often overwhelming for the clinical staff, burdening their decision making processes. With recent advances in the field of machine learning (ML), there has been growing interest in exploiting the wealth of ICU data for tasks such as prediction and decision support [1–5].

Current work in ML for the ICU mostly focuses on two types of tasks: The first is predicting a deterioration in a patient's health, commonly implemented as an alarm notifying the clinical staff before a critical event [6,3,7,4,8]. This task is considered particularly interesting due to its potential for improving patient care and saving lives. It was also reported as the ML use-case ICU clinicians are most likely to adopt [9]. The second prominent line of work applying ML in the ICU setting focuses on predicting patient vital signs such as heart rate, respiration, blood pressure and more. Such vital signs are considered to be the main window onto what clinicians generally denote as "patient state" [1] [10–16]. Unfortunately, recent attempts for deployment of such prediction systems did not lead to better outcomes in any of the reviewed performance measures [17,18]. It seems these failures are not strictly caused by algorithmic malfunctions, but result from operational and human constraints [17]. However, despite the extensive research and substantial resources invested in the task of real-time ICU prediction (over 970 articles and counting [6]), to the best of our knowledge no prior work investigated clinicians' requirements from ICU prediction models.

Therefore, in order to better inform the design of ML-based prediction tools that fit clinicians' workflow in the ICU, we conducted a mixed methods study that included both open-ended interviews and quantitative tasks with ICU staff across four ICU wards in three different hospitals. We held exploratory interviews focused on clinicians' needs in the context of ICU prediction systems. These were followed by quantitative tasks aimed at evaluating the perceived clinical relevance of predictions that could be obtained from machine learning algorithms. Our study revealed several insights that can facilitate the design of clinically useful ICU prediction systems. In particular, some of our findings suggest approaches that differ from current practice.

* Corresponding author.
E-mail address: briany202@gmail.com (B. Eini-Porat).

[1] We note that *patient state* is a widely used yet loosely defined term, referring to the underlying patient health, often assessed by clinicians based on the patient's vital signs.

https://doi.org/10.1016/j.jbi.2022.104107
Received 28 February 2022; Received in revised form 26 May 2022; Accepted 28 May 2022
Available online 7 June 2022

Table 1
Participant details, including years of experience.

ID	Role	Experience (Years)	ICU type
P1	Physician	5	Adult
P2	Physician	30	Adult
P3	Physician	17	Adult
P4	Physician	1	Pediatric
P5	Physician	2	Pediatric
P6	Physician	25	Adult
P7	Physician	1	Adult
P8	Physician	44	Pediatric
P9	Physician	20	Pediatric
N1	Nurse	11	Pediatric
N2	Nurse	16	Pediatric
N3	Nurse	10	Adult
N4	Nurse	10	Pediatric

2. Related work

Due to the challenges in uptake and sustained use of ML systems in healthcare [19], previous work has investigated the needs of their clinician end-users [42]. Tonekaboni et al. [20] conducted a target stakeholder study for the purpose of identifying explainability challenges in ML for healthcare in general, and proposed strategies for enhancing trust by focusing on clinicians' explainability needs. Some studies focus on user-centered design meant to support specific patient care scenarios [21], and others strive to co-design AI based decision support tools involving clinicians [22]. Yet, none of these works focus on ML-based prediction tools for the ICU setting.

More generally, despite the extensive research and substantial resources invested in the real-time ICU prediction task (over 970 articles [6]), to the best of our knowledge, no prior works have investigated clinicians' requirements from ICU prediction models. Some pilot deployments of such prediction systems to the ICU or emergency department were extensively documented and include deployment pain points such as the need to adapt to current workflows [23,24], or evaluated their usability post-hoc[43].

Clinicians needs in terms of current monitoring (as opposed to future predictions) and workflows use of information in the ICU are the subject of multiple studies [25–27,9,28,41]. Among their findings are the intuitiveness of vital signs monitoring, as opposed to the difficulty of tracking other kinds of information. Moreover, alarm fatigue was identified as one of the major threats to patient safety. However, these works focus on deployment and interface design for monitoring and information systems, rather than ML-based *prediction* tools.

A survey study by Poncette et al. [27] focused on the satisfaction of ICU staff with current patient monitoring and suggestions for future improvements, and included questions regarding future incorporation of AI-driven tools. Among several possible use cases for AI in the ICU, participants indicated they would use AI-based tools to predict complications and detect increased risk of mortality. This paper presents areas of focus for development of AI-based tools, but does not specify any algorithmic requirements.

While we believe interface and explainability requirements are crucial for successful deployment, we also believe that functional requirements must be addressed. Therefore, our work considers the underlying algorithms as a crucial part of the design space, particularly focusing on the ICU setting, in a way that to the best of our knowledge has not yet been addressed by prior work.

3. Methods

To gain a better understanding of the challenges clinicians face in the ICU and the potential of machine learning systems to support their work, we conducted a mixed-methods study involving clinicians from four ICU wards across three hospitals. The study comprised of semi-structured interviews and quantitative tasks. The interviews were designed to understand clinicians' thought processes and elicit their considerations, in order to asses their requirements from ML tools predicting patients' health state in the ICU. We further asked the participants to complete tasks designed to evaluate specific quantitative aspects of the clinical relevance of predictions that could be obtained from machine learning algorithms. The study was approved by the Technion institutional review board.

3.1. Participants and recruitment

In the course of this study we worked with ICU clinicians including physicians and nurses from both adult and pediatric ICUs in 3 different hospitals. Our interviewees represented a diverse sample including nurses, interns, residents, fellows, attending senior physicians and two departments chiefs; their mean work experience was 15 years with a range deviation of 1 to 44 years; see Table 1. All of the participants completed both parts of the study. Participants self-reported a mean level of familiarity with statistics (3 ± 0.9) and machine learning (2 ± 0.6) on a 5-point Likert scale (*1 - Heard of it* and *5 - Expert*). We recruited participants using a snowball sampling approach, starting from clinicians in the hospital in which one of the authors practices as a clinician. The recruitment of potential interviewees continued until we have reached saturation of new concepts in the exploratory interviews, meaning no new concepts were revealed in the last three interviews. Potential participants received information about the study and its requirements. Participants were compensated with coffee vouchers (worth $\sim$ 8$) to thank them for their time.

3.2. Data collection

Interviews were conducted by the first author in one-on-one sessions, with duration ranging from 40 to 100 min depending on participant cooperation and availability. The interviews were recorded and transcribed following participant consent, with the exception of three interviews which were not recorded due to participants' refusal; these were documented using notes. Two of the interviews were conducted remotely due to Covid-19 restrictions. A fundamental principle of the study design in both its qualitative and quantitative parts was starting with open-ended questions and moving toward increasingly structured questions as the sessions progressed. We made this choice in order to

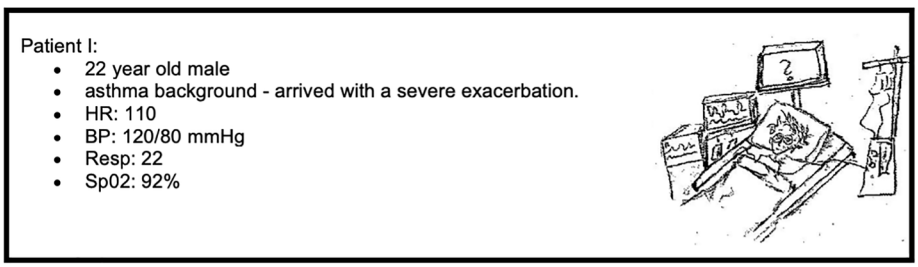


Fig. 1. Example patient case. This case is considered less severe in the ICU.

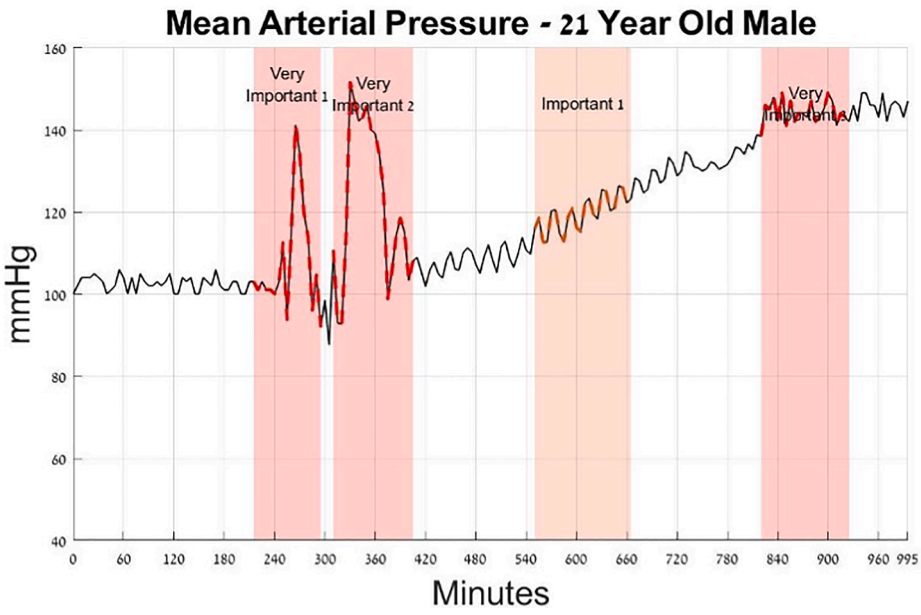


Fig. 2. Example of an importance heat map over blood pressure values given by one of the participants. Participants were provided with such trajectories and context and were asked to mark *important* or *very important* segments. In this figure, three segments were marked as *very important* and one was marked as *important* by the participant.

reduce the possibility of introducing our own biases with leading questions.

3.2.1. Exploratory interviews

Prior to the interviews, we generated 4 scenarios of ICU cases depicting standard ICU admissions with varying severity (cases are detailed in A). Participants were presented with 2–3 randomly selected ICU patient cases. Patient cases included basic information: presenting symptoms, age, gender, vital signs (heart rate, blood pressure, SpO2, respiratory rate) and an illustration; see Fig. 1. In order to allow both pediatric and adult intensivists to participate and contemplate familiar scenarios, patient cases focused on young adults that are treated by both clinician groups.

The purpose of the presented cases was to ground the discussion on current workflows, identify important considerations and ease the transition towards a discussion of envisioned prediction systems. To guide the semi-structured interview, participants were asked the following questions per case: *(1) What would you check about this patient? (2) What are the next steps?* After discussing current workflows, we introduced the notion of prediction systems in this context: *(3) Say we have a system that can make predictions with respect to this patient, what would be helpful to know? why? (4) What time frame should be used for a prediction update?* If the subject was not brought up naturally in the conversation we also asked *(5) Would you prefer predictions in the form of future events or future patient state?* We continued to discuss the integration of the desired predictions into current ICU workflows.

We developed the initial interview guide with the goal of understanding clinicians' needs from prediction systems, as well as assessing the compatibility of these needs with current practices for ML-based prediction model development. To allow perceived needs to spontaneously come up in the conversation, we kept our questions open-ended. We conducted two pilot interviews following which the interview guide was refined: the number of patient cases presented was reduced, and question (5) was added to complement question (3).

3.2.2. Quantitative study protocol

After completing the semi-structured interviews, participants preformed a series of four tasks designed to measure specific aspects of perceived utility from observing different vital signs predictions. The decision to focus on vital signs for the quantitative part was made a priori, as clinicians tend to describe and track patients' state using vital signs. Furthermore, any reasonable ML model used in a real-time capacity in the ICU setting is likely to use vital signs as an important building block. The tasks were presented in the following order:

T1: Identifying important events The quantitative session started with a relatively neutral think-aloud like task [29] which allows the subjects to express their thoughts freely. In this task, the participants were asked to identify (and if possible rank) important events and/or time intervals in sample vital sign trajectories presented to them, and mark them as *important* or *very important* using a simple interface. Specifically, they received the following instructions - *Mark segments of this signal describing predicted events you be made aware of or be informed about while caring for a patient (if such exist). You may choose to mark them as important or very important if you can. Please guide us through your thought process and verbalize your thoughts.*

We used vital sign data from the MIMIC-III Wavefrom Database [30], which contains thousands of vital signs time series collected from bedside patient monitors in intensive care units (ICUs). The sampled vital signs were aggregated at 5 min time steps and reflect the mean value within the 5-min interval. This aggregation enables the inspection of long trajectories (hours) during which interesting events are likely to occur.

Specifically, participants were presented with at least four randomly sampled vital signs trajectories of adult patients from the MIMIC-III Wavefrom Database, infused with a few synthetic events (including exceeding normal range and sudden spikes), while making sure they mostly follow the original data; see Fig. 2 for an example. The vital signs we used were the trajectories of mean atrial blood pressure (MAP), heart rate (HR), arterial oxygen saturation (SpO2) and respiratory rate (RESP). For each vital sign, the patient's age and gender were presented to provide context. Participants were told that these patients were admitted to the ICU and are currently in a hospital bed, but with no further context. We note that in the ICU, clinicians would have at their disposal additional information and more specific context such as patient comorbidities, which might change their actual preferences. However, accounting for multiple possible contexts, even brief patient histories as in *T1*, would greatly increase the variety of patient cases we must present to the participants, which in turn would have required a

20 year old male, no chronic conditions

HR

BPM	30<	30-40	40-50	50-60	60-70	70-85	85-90	90-100	100-110	110-120
Rank										

BPM	120-130	130-140	140-150	150-160	160-170	170-180	180-190	190-200	200<
Rank									

Fig. 3. Heart rate interval scoring example. Participants were presented with the intervals above and received the following instructions: *Please score on a scale of 1–10 the significance of knowing the next vital measurement for a measurement within the following intervals.*

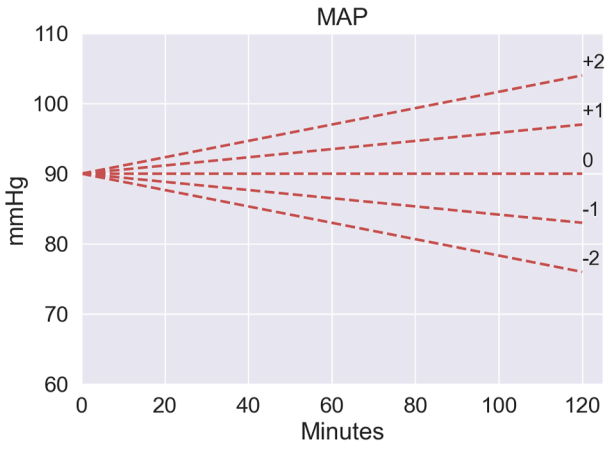


Fig. 4. Example mean blood pressure trajectory. The trajectory is described over the course of two hours without exceeding normal range. The number next to each trajectory refers to its slope – participants were presented with the trajectories without this slope notation.

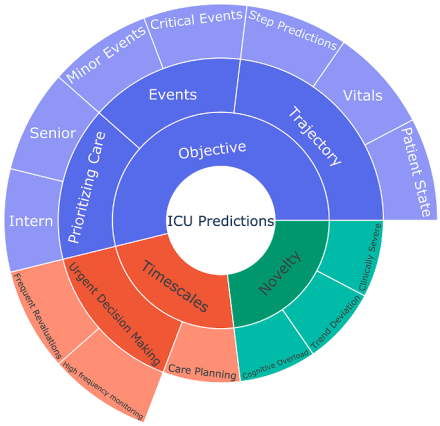


Fig. 5. The core themes that emerged from the study. Within three categories (inner ring), 8 higher-level themes (middle ring) and 9 sub-themes were specified (outer ring) to identify the algorithmic requirements from ICU prediction models in the view of intensive care staff. Every category has a different color, with paler shades for the lower level themes.

great amount of time resources from the clinicians involved, exceeding their limited availability for our study.

Depending on the quality of verbalization, participants were asked questions regarding their actions and motives. These probing questions were asked during the execution as a form of coaching - reminding participants to verbalize their thoughts when they became silent: *Why did you mark this segment? Is this segment more important than the previous one? Why?*. This task preceded the following scoring task in order to avoid bias and to encourage original thought.

T2: Interval Scoring The next task required scoring the significance of observing certain prediction values of vitals for a hypothetical patient. This task was designed to derive a quantitative formulation of the clinical relevance of the absolute nominal values of predicted vitals. This task was preformed with respect to a certain patient, so the evaluation relates to a more familiar use case. The scenario and vitals matched the previous task. Specifically, participants were asked to score intervals of hypothetically predicted vital signs values according to the attention required for monitoring the patient, see Fig. 3. Intervals were displayed together to encourage relative comparisons of importance of the information. After each scoring task the subjects were asked questions regarding their scores (*Why did you score X higher than Y?).*

T3: Anchoring This task is identical to the Interval Scoring task T2 with one key modification – displaying a previous measurement along with the patient's basic information. The anchoring task captures a certain aspect of context. It reflects on the change in clinical relevance attributed to a prediction given additional information regarding previous measurements. Two versions of this task were used: one displaying a measurement from the previous hour and another displaying two measurements, from one hour and from 30 min prior. Each of the two versions was completed by half of the participants. We note that the two measurements variation was added after the first five interviews, following early results. Its purpose was to evaluate deviations from trends.

T4: Pure Trend The pure trend task aimed to measure the importance of a trajectory's trend separately from the absolute values of the vital sign. Participants were presented with two hour trajectories of the four vital signs (HR, MAP, SpO2 and RESP) and asked to score the level of importance they ascribe to them; see Fig. 4. In order to disentangle the effect of trend from that of nominal values, trajectories remained within each vital's normal range. We note this task was also added after the first five interviews, following early results.

3.3. Data analysis

We analyzed the data using a thematic approach. The interview data was coded using open coding analysis [31]. As a first step, the first author reviewed three transcripts to generate an initial coding scheme. It was then discussed and revised by the entire research team who reviewed suggested codes with the corresponding examples of participant responses. Then, the remainder of the interviews were coded, allowing for new codes to be developed, reviewed and finalised by the team (e.g. "predict trend", "Abnormal attributes", see E). An external coder was invited to code a third of the transcripts, initially coding two transcripts independently. The coders discussed the results, discussing

the few disagreements that emerged, which were mostly related to the level of granularity of the codes. Then, they proceeded to code the remaining transcripts according to the coding scheme and reviewed the codes together yielding inter-rater reliability of 82%. The few disagreements were resolved by discussion. Further analyses were done using affinity diagramming [32], whereby the lead author began merging codes into initial themes. All authors iteratively clustered data into themes (e.g. "vital signs trajectory", "high frequency predictions"). The emerging themes were reviewed and revised over the course of several sessions. These themes had been sorted to create higher-level themes or categories, see Fig. 5. Thus, at the end of this process, we had identified design implications as well as the specific themes associated with each. We further reviewed participants' responses to the quantitative tasks and plotted them in aggregation to identify recurring themes.

4. Results

4.1. Findings from the exploratory interviews

The complexity of the ICU environment and its associated diagnostic and treatment challenges are apparent from the exploratory interviews. Several core concepts and insights emerged consistently in the interviews; we focus here on those that pertain to predictions in the ICU settings and are novel.

4.1.1. Critical events vs. patient's trajectory predictions

Many participants emphasized that while their main goal remains averting critical events and improving treatment outcomes, they would rather have predictions regarding patients' future vital signs trajectory (representing the patient physiological state) or general patient state, and use these to guide clinical decisions. Moreover, the specific vital sign value appears to be less important than the change/trend from the previous step or even the patients' trajectory: *"I want to see change in pulse – I see it going up 112, 116, 120, 130, 150. A system should monitor difference, trend" -P9*. The clinical relevance of the trajectory should also be taken into account as it is often discussed in this context: *"Would like to see a physiological-well-being graph" -P2*. Naturally, deterioration was mentioned more often than recovery and is more important to identify: *"I don't care as much what would be the rate of the recovery" -P1*.

Participants also noted that subtle trends are more likely to go unnoticed, especially if the ward is busy, and thus trajectory prediction could make these changes more apparent and relieve some of the cognitive load associated with patient state tracking. Moreover, participants also expressed their interest in counterfactual predictions, given specific alternative interventions in the same abstracted context of trend rather than the specific value. Trajectory prediction was raised by some of the clinicians as a means to allows less aggressive interventions at an earlier stage, where the error cost of unnecessary intervention is low. For instance, participants noted that *"if the patient looks fine at the moment, but gets a bad prediction, many won't agree to treat aggressively now. I can't send this patient to surgery" -N4*, or that they *"Can just administer antibiotics.. preventive medicine is best." -P5*.

4.1.2. Prediction timescales

Any model for continuous prediction has it own update rate or prediction horizon timescale. Ideally, predicting according to the progression rate of a specific clinical condition would allow appropriate decision support. However, there are countless conditions and diseases; furthermore, multiple conditions can manifest in a single patient. Nevertheless, our study participants identified two main prediction timescales, each with a different role. The first timescale is 1–3 days and is meant for high-level planning of care. The second timescale is less than one hour, and is meant to support immediate and urgent decision making for unstable patients in need of acute care. These timescales were revealed either by discussing different patient cases or as a specific requirement: *" I would like this long-term (3 days) prediction. In the ICU we don't look so far ahead, but in medicine, prevention is most effective for saving lives […] In context of acute medicine, of course I would like short terms predictions" -P5, "It would be helpful to get predictions in a 15 min interval because it's difficult to notice subtle changes" -P6*. The state of unstable patients in acute care rapidly changes and requires constant decision making. According to the more experienced clinicians, frequent re-evaluations and decision points are considered as best practice: *"Specifically in the ICU, we make decisions all the time, the best practice is to keep the time between decision points as short as possible. It is possible I have a certain impression about a patient and re-evaluate minutes later. That way you can fix easily" -P8*.

4.1.3. Prioritizing care

Although we presented prediction systems in this study in the context of a single patient, participants repeatedly emphasized that a main added value of such prediction systems would be aiding them in prioritizing care. Clinicians seem to be quite confident in their ability to provide an adequate level of care for a single patient. However, in large or busy ICU wards, or during night shifts when clinical staffing is shorter and includes less experienced staff, the cognitive workload of each clinician increases. It might be easier to miss a deterioration in patient's condition: *"One of the main issues in large ICUs is prioritizing […] I can deal with a single patient […] But if I am treating this one patient that looked bad, it is very easy to get sucked into the treatment and miss another patient […] At 2:00AM there is no one else that sees all the beds at once. Who is getting better and who is getting worse? When I am focused on a single patient I might be biased" -P5, "If I have 200 patients, I would like to know which are the top-10 patients that we need to focus on the most […] If no one is looking at the patient in room 7 then they might not notice the change until it is too late" -P2*. These claims were made both by experienced clinicians and by residents.

4.1.4. Tell me something I don't know

Participants have also mentioned the significant cognitive overload caused by the multiple monitoring systems already in place and their corresponding alarms. Any system that is deployed in an ICU setting takes its cognitive toll and should therefore have a clear added value in terms of novelty and actionability: *"Our staff doesn't like too many systems. Each added system is another screen to monitor, another alarm to listen for. If you add a system, it should have practical added value – tell you something you don' know and can act accordingly" -N3*. This element of novelty or surprise also refers specifically to trajectory predictions: *"I can see where this patient is going and understand that it [the trajectory] deviates from the expected trajectory" -P6*. While some participants prefer having an abundance of information with respect to the prediction, including explanations (possible diagnosis, event probabilities), others argue that predictions should only indicate that something is wrong, encouraging the clinicians to look for a cause for deterioration and possibly facilitate the detection of human errors. They further claim detailed predictions would hinder the clinical staff, especially if the explanations provided are not complete.

4.2. Estimating clinical utility

In this section, we look into the quantitative aspects of perceived clinical utility of vital sign prediction. Investigating the utility is especially relevant given the clinician's reduced reported interest in prediction of only absolute measurements: clinicians seem more interested in the overall trajectory and patient trend according to their perception of severity rather than the momentary values of the vital sign. In their view, clinical utility is directly related to the perceived clinical severity (Examples in B); we wish to quantify this utility in the following tasks.

When presented with examples of predicted vital sign trajectories, clinicians consistently marked areas that they perceived as reflecting two aspects: clinical severity, and surprise, with both aspects depending

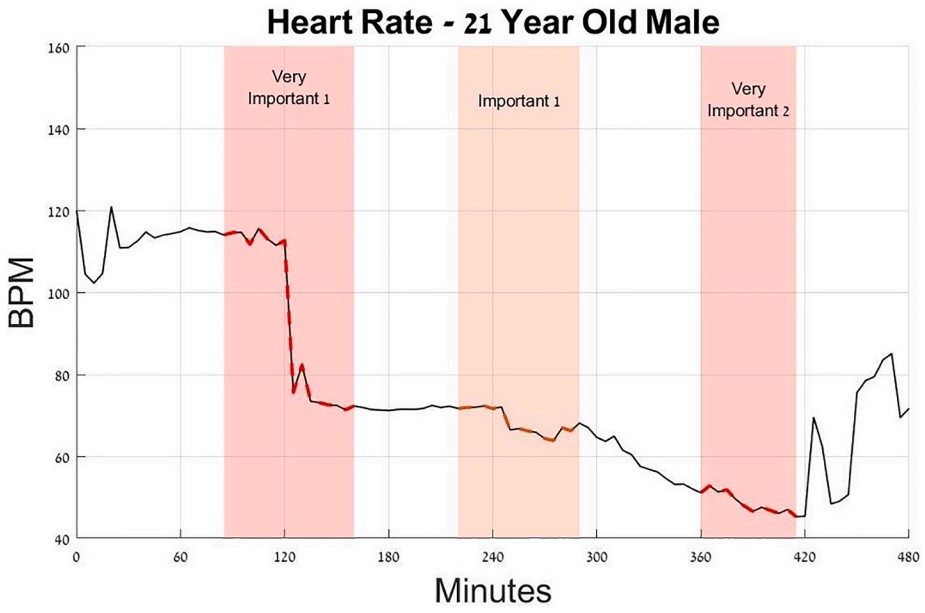


Fig. 6. Task T1. Example of an importance heat map over heart rate values given by one of the participants, including the three main components for vital prediction: deviation from clinical norm, overall trend, and surprising deviations from said trend.

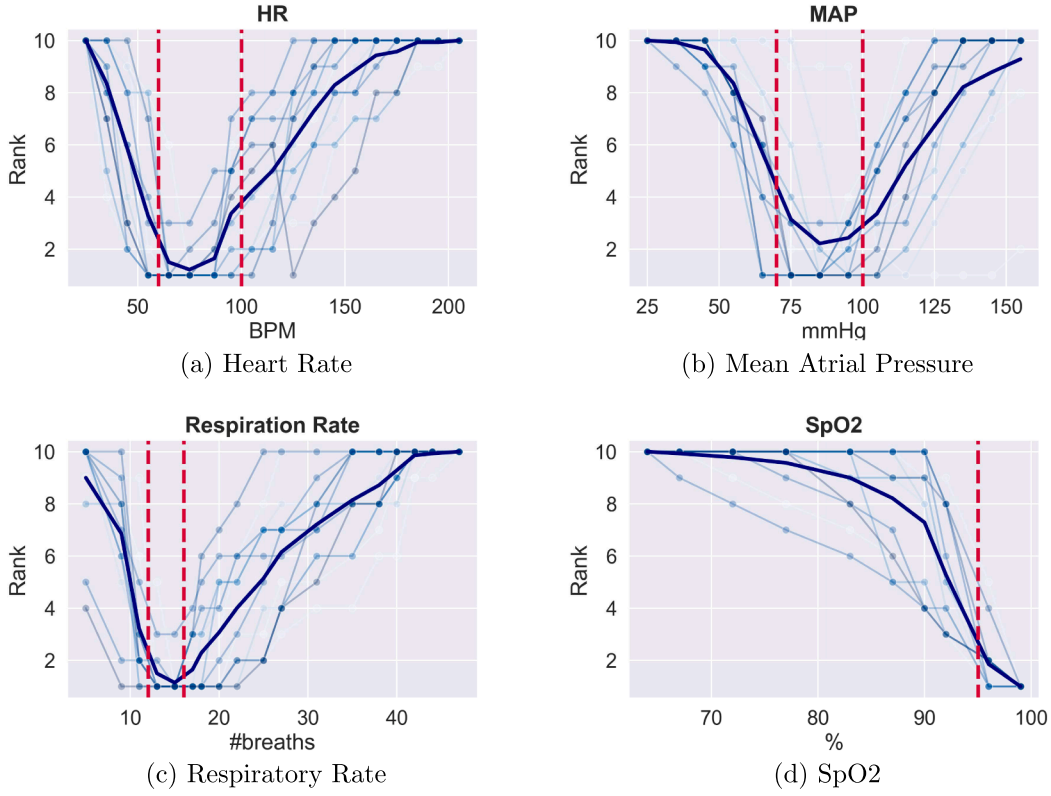


Fig. 7. Clinical importance of absolute values per vital sign. The widely-accepted clinical normal range is marked in red dashed line. Mean values are in dark blue, while light blue lines show the rankings of individual participants.

on context. Clinical severity is determined by the absolute values and the trend of the presented vital sign. Surprise depends on the deviation from an expected trend given previous values.

Participants often mentioned the difficulty of assessing the true importance of predictions in specific areas without further context, as context could affect the expected behavior. They also sometimes struggled to differentiate between 'important' and 'very important' segments. Nevertheless, three components of signal behavior were consistently identified as important for all signals: general trend, sudden deviations from the trend and severity of absolute values as compared to the normal ranges. Fig. 2 shows an example from one of the participants: this participant marked as important sudden spikes in blood pressure, rising of blood pressure, and steady abnormal values (for more examples see B). We note that participants' tolerance to deviations from the normal

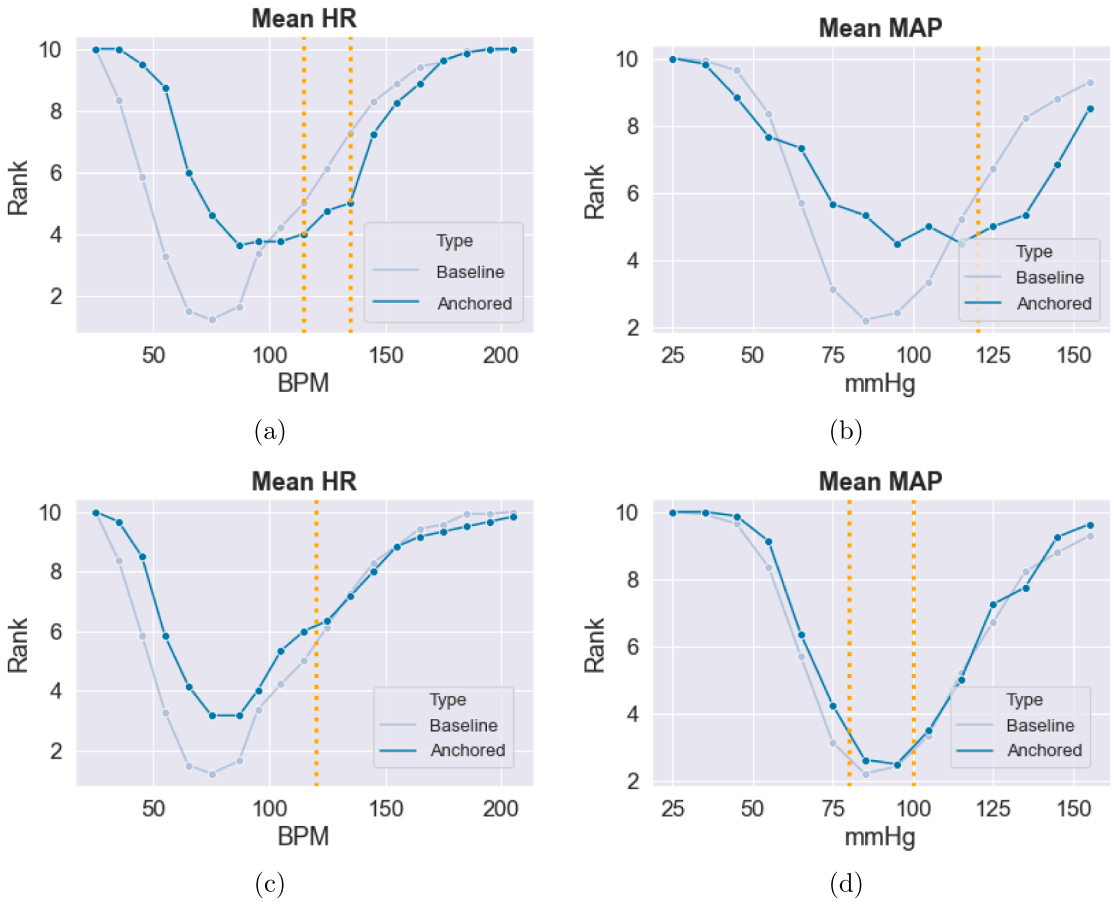


Fig. 8. Plots (a) and (b) depict scenarios in which an anchor (orange dashed line) is located beyond the normal range in which participants expressed concern. In plot (d) anchors are within normal range and in plot (c) they are borderline. In plots (a) and (b), the curves are importance elevated but also wider. In plot (c), the curve is elevated but similar to baseline. In plot (d) it is similar to the baseline.

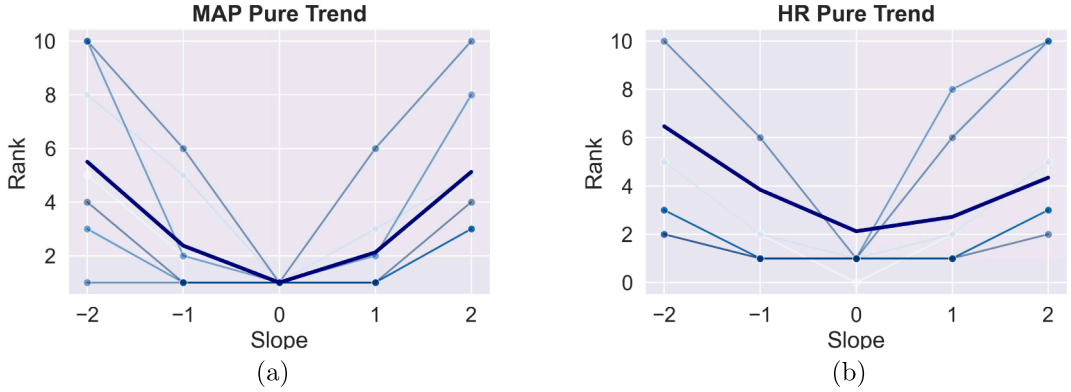


Fig. 9. Ranking of importance of mean atrial pressure and heart rate trajectory slopes. Similarly to Fig. 7, mean values are in dark blue, while light blue lines show the rankings of individual participants.

range varied across signals: participants were willing to tolerate a wider range of values and steeper changes when it came to respiration rate compared to blood pressure and heart rate.

When asked to explain their choices, participants provided different takes on clinical severity. High or low absolute values are important as they describe clinically severe states - a patient with a predicted heart rate of 180 BPM is probably in danger. General trend hints at clinical severity even if the absolute values are not severe just yet – the clinical staff could be missing a deterioration. As for sudden deviations from the trend, participants provided two different explanations for their importance: One is that a sudden change could be a symptom of a clinical event. The other is the element of surprise, which by definition is less likely to be predicted by the clinical staff and therefore the prediction of a surprising event gives them important information. In addition, the importance of both general trend and deviations from it clearly interact with values exceeding the normal range. For example, the same slope of heart rate trajectory is more important when the absolute values are also concerning. An example of this is shown in Fig. 6: here, the participant marked the same trend slope as more important as it progressed toward bad absolute values.

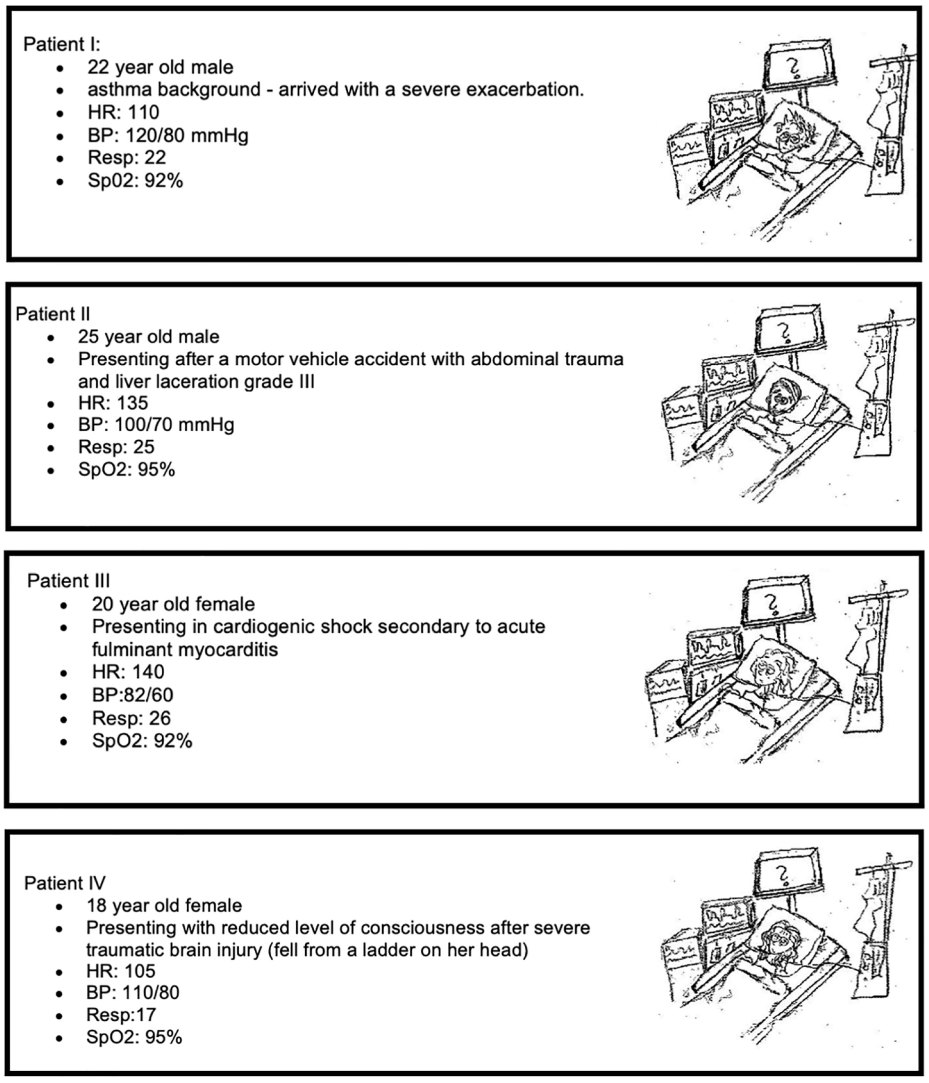
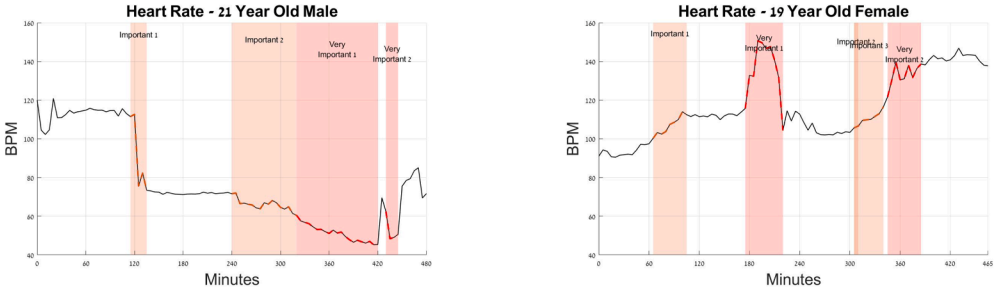


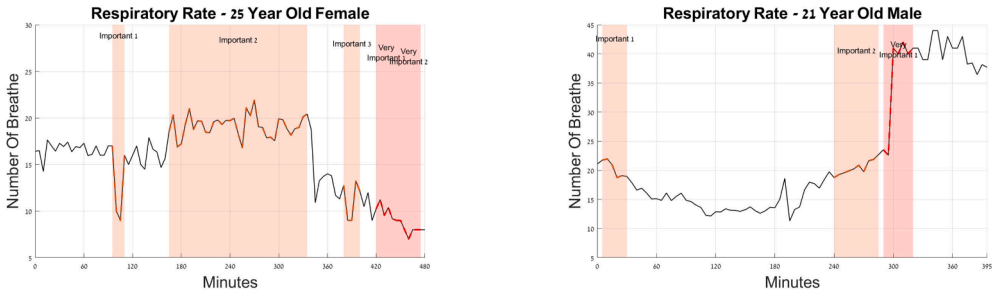
Fig. A.10. For each of the patient case scenarios additional context was presented, including age, gender, and basic vital signs.



(a) (b)

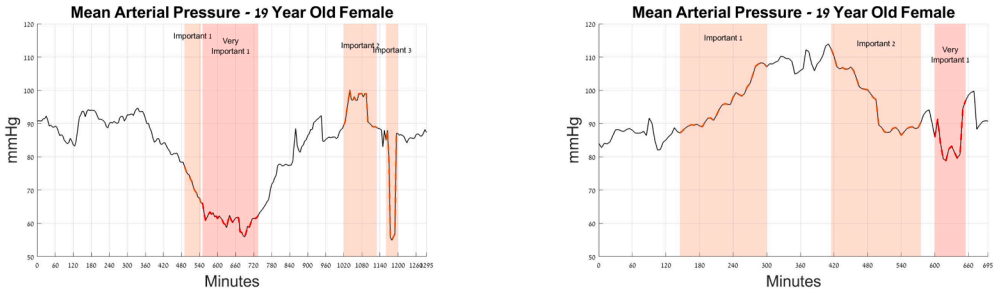
Fig. B.11. Examples of heart rate importance heatmaps marked by participants. blue (a) reflects sudden changes and interaction between trend and values. In (b) we observe the importance of trend, with a sudden peak marked as a very important event, and again trend is considered worse when we get to clinically severe values.

In Fig. 7 we can see the clinicians' ranking of absolute values according to their clinical importance; the rankings yield a collection of asymmetric curves with sharper slopes in areas where the signal exceeds or falls short of clinically acceptable normal range. For example, when observing participants' rankings for respiration rate values, we can see that values within the clinical normal accepted range (12–15 breaths) receive low importance. When exceeding normal range toward low values (signaling an impending respiratory failure), importance increase rapidly, yet when exceeding normal range toward high values (signaling increased respiratory distress, but not failure) the importance curve increases at a slower pace. For each direction of exceeding norm, it appears as an S-shaped curve, with moderate slope in the normal range interval, then a steep rise, and moderate slope again. Note that SpO2 can only fall below the normal range, whereas the three other signals can



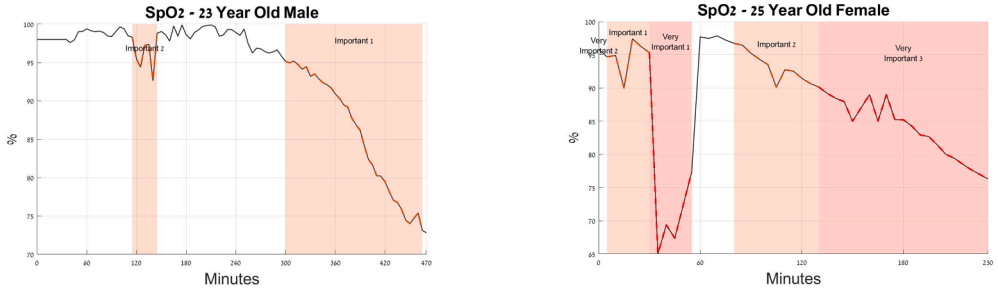
(c) (d)

Fig. B.12. Examples of respiratory rate importance heatmaps marked by participants. (c) Shows that elevated values are important, final trend is worse due to the absolute values. (d) Depicts the importance of sudden change and its amplitude.



(e) (f)

Fig. B.13. Examples of mean atrial pressure importance heatmaps marked by participants. It is apparent from (e) that the participant tolerates severe values given the patient history. (f) Depicts the importance of trend vs. sudden deviation from trend.



(g) (h)

Fig. B.14. Task T1: Examples of O2 sturation importance heatmaps marked by participants. (g) and (h) are an example of the variability in marking.

reach values both above and below normal.

We note that participants expressed concern only when values are well beyond the normal range threshold, suggesting that the range of acceptable values in the ICU is wider than the normal range as it is currently defined in the medical literature.

However, the exact threshold of concern seems to change according to previous measurements "anchoring" the observations: the results are presented in Fig. 8: Previous measurements that exceed normal range result in a greater baseline concern for prediction in the next step, even when it manifests within the norm. However, participants tolerated a wider range values when they were closer to the previous values. Previous measurements or trend within normal range do not change the curve. For example, in the MAP signal, when the previous measurement is 120 mmHg (Fig. 8(b)), which is well beyond clinical norm and within the value range that is considered concerning in task *T2* (see Fig. 7), the curve is elevated and wider than the baseline. But when previous measurements are within clinical norm the curve is identical to the baseline (Fig. 8(c)). In general, having a previous measurement within clinical norm resulted in the same curve as in *T2*; for more examples see C.

When examining participants' view of the trend, it seems there is a non-linear increase in importance as vitals' slope becomes steeper. For example when dealing with MAP (Fig. 4(a)), the difference between slope -2 to -1 could not be described by the same straight line as the difference between slope -1 to 0. The exact behavior is slightly different across signals, but the general shape remains the same; see also Fig. D.16 in the Appendix. However, there is some notable variability in ranking among participants. While none of the participants expressed concern about the "no change" option, participants were divided in their responses to steeper trajectories. For each signal, about half of the participants assigned them a high importance score and the other half gave low-to-middle scores. This phenomenon occurred across experience level and profession. Furthermore, even participants who reported they would prefer predictions of patient trend assigned such lower scores occasionally. To conclude, it seems that pure trend is important by itself,

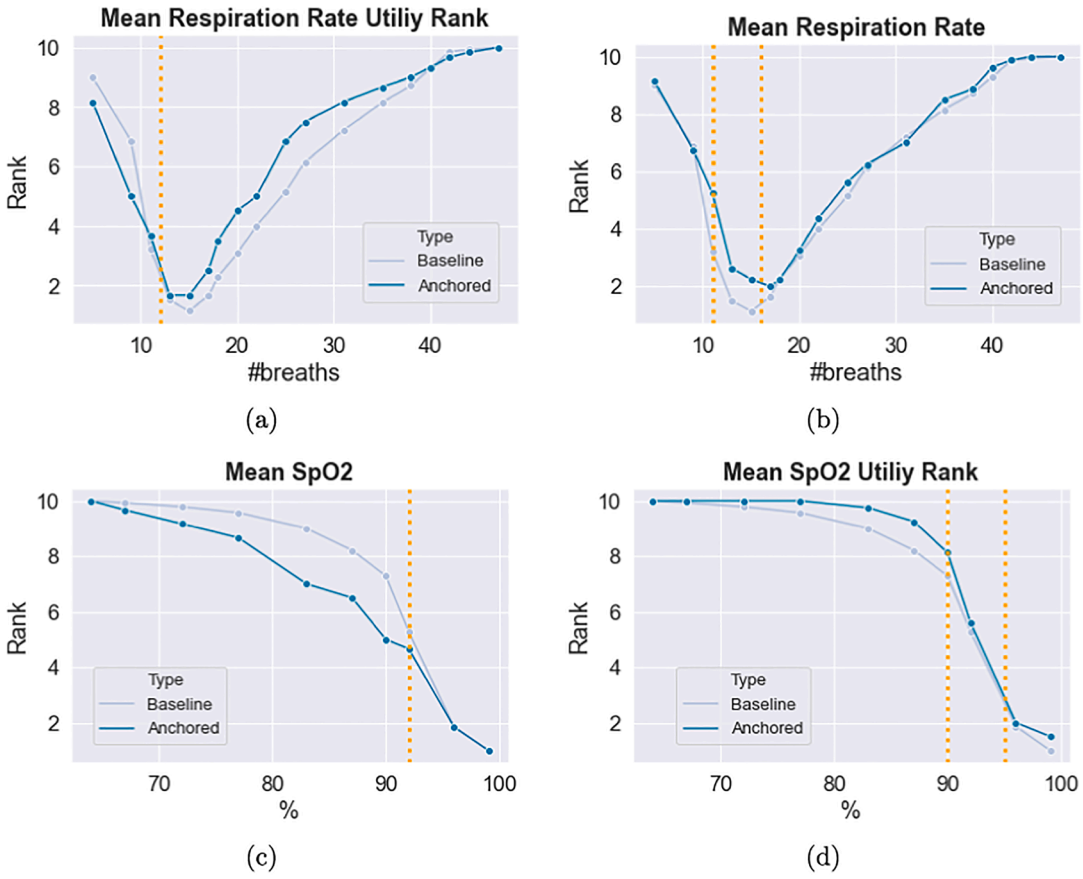


Fig. C.15. Change in importance ranking when previous measurements are within clinical norm. SpO2 previous measurements are borderline, deviating from clinical norm but remaining in intervals of medium concern as measured in task *T2*. Figure (d) is slightly more severe than (c) as it depicts a steeper deterioration.

Fig. D.16. Ranking of importance of mean atrial pressure and heart rate trajectory slopes. Mean values are in dark blue, while light blue lines show the rankings of individual participants.

but less than absolute values. (See Fig. 9).

5. Discussion

Our work has revealed clinician preferences regrading predictions from ML-based systems, highlighting a need for predicting patient trajectories and assistance in prioritizing patient care. The participants posed two possible aims for predictions systems in the ICU, corresponding to two different timescales: a days-long timescale for high-level care planning, and a short (< 1 hour) timescale for frequent re-evaluations of care. In all cases participants emphasized the importance of novel, actionable, predictions over successful predictions of non-surprising occurrences. Finally, quantitative tasks revealed a non-linear utility function for different predictions, valuing predictions of deviations from clinical norm, overall trend, and deviations from said trend.

5.1. Design implications and requirements

Predicting patient trajectories rather than critical events. While the majority of deployment attempts of ML-based prediction systems to the ICU focus on predicting specific critical events (as an alarm system) [33,17], many of the study's participants showed a preference towards displaying predictions differently. It was clear from the interviews that the prediction of stand-alone clinical events remains a source of interest. Yet, our study also suggests possible difficulties in clinicians' perceived ability to act preemptively and decisively based on such predictions alone (see 4.1.1). Since predictions of acute events may never materialize, some clinicians would rather defer aggressive interventions in response to what they view as uncertain, possibly black-box, predictions. Notably, several clinicians stated that they see utility in building models predicting interventions (as proposed by Suresh et al. [34]), with a preference for predicting less aggressive interventions such as antibiotics administration or other routine procedures. Clinicians indicated that they are more likely to act upon predictions when they consider the derived actions to be of relatively low risk. This preference for predictions with low-stakes implications might indicate a lack of trust in clinicians' perception of the possibilities of prediction systems in general, as well as a reluctance to engage with highly uncertain predictions.

An intriguing observation is that clinicians seem to be in interested in more than just events. As described in Section 4.1.1, many prefer to observe a prediction for the patients' future trend, i.e. prefer predictions of the future trajectory over single point predictions. Such prediction models were proposed for example by Clifton et al. [11], Schulam and Saria [12], Colopy et al. [15], Alaa and van der Schaar [35], Cheng et al. [16]. Displaying continuous predictions rather than just alerting about upcoming critical events may uncover subtle changes in patients' health and raise the clinicians awareness to such fluctuations. This in turn allows for preemptive, repeated clinical evaluations and proactive, less aggressive interventions, compared to those needed in case of a critical deterioration event. Therefore, the preference for trajectory prediction seems to again reflect a preference for a system enabling subtler clinical interventions, in line with the preferences stated above for predicting milder, yet more common, events compared to prediction of rare, acute events. Towards this end, we believe that displaying future trajectory information without the need to read specific values might ease the processing of this additional information and enable proactive action. We also note that current work in medical informatics aims to relieve cognitive load in clinical decision-making by displaying patient data and organizing it into easy to understand temporal resolutions [36,37]. Displaying trajectories aligns with this approach and allows better understanding of patient state as an evolving process.

Model Evaluation and Optimization within Context. A prominent issue that repeatedly came up both in the qualitative and quantitative interviews is that the utility and interest in the predictions is highly context-dependent. This is true since even patients who are severe enough to be admitted to the ICU remain relatively stable most of the time. Detection of a sudden drop in heart rate is more important than correctly predicting "no change" multiple times. Thus, standard evaluation metrics provide an insufficient way of assessing the utility of the prediction for actual clinical practice.

What emerges from the interviews is that the perceived importance of a prediction depends on a combination of three factors: (i) deviation from clinical norm, (ii) overall trend, and (iii) surprising deviations from said trend.

For deviation from clinical norm, Task *T2* (subSection 3.2.2) yielded the importance curves in Fig. 7 which clearly demonstrate that the difference between values is not linear, whereas standard evaluation metrics such as squared loss treat them as equal. These curves could be directly utilized in the construction of new evaluation metrics designed to account for actual clinical utility. We note that the curves align with the notion of subjective expected utility from classic decision theory [38]. The expected utility at a certain point depends on a subjective probability weight which determines whether an event will occur: $w(y_t) \cdot u$, where the factors u and w correspond to the cost of an adverse event and the subjective weights model to the *amount of attention* given to each event y_t, respectively. Therefore, drawing on this theory in conjunction with empirical studies can further support the implementation of more informative evaluation metrics.

Tasks *T3* and *T4* (Figs. 2 and 8 and B and C) reflect the weight given by clinicians to the relation between consecutive values, thus indicating the weight of overall trend (ii) and surprise (iii) in the clinical utility of predictions. Again, trend is mostly disregarded by current evaluation metrics, but in reality, previous observations create certain expectations regarding consequent measurements. Trivial predictions (e.g. "no change") seem to be less interesting even if they indicate a clinically severe state. Furthermore, a steep trend or a sudden trend deviation in vital signs often imply a change in the patient's general health. The curves in Figs. 2 and 8 can be used to formulate new evaluation measures that capture the clinical utility of a prediction with respect to previous observations and the overall vital sign trend.

Prediction Timescales. Our findings suggest two distinct prediction timescales (see 4.1.2), each serving a different purpose: One is a 1–3 day timescale, and the second is less than an hour. The former is meant to support high-level care planning, while the latter is meant to support frequent re-evaluations and decision points required in the ICU. Since ICU prediction algorithms are designed as decision support instruments, they should comply with the current rate of decision making. While some works do predict for the slow timescale [6], to the best of our knowledge, there are only a few works that predict for less than an hour [39,11,40], highlighting a currently unmet need in the field.

Prioritization of Care. Finally, our study reveals a possible avenue for using ML to improve ICU patient care that differs from most of current practice: instead of optimizing predictions for a single patient, solving an attention prioritization problem (subSection 4.1.3). Predicting for the single patient does not account for an ICU's limited resources, where clinical staff typically must attend to multiple patients at the same time. Night shifts or extremely busy shifts are the most vulnerable situations; their common pain point being the challenge of "who to attend to first?". We note that any algorithms addressing the prioritization problem will need to take data of the entire ward as input rather than a single patient's data. However, prioritization in the ICU is not merely based on clinical severity, but on actionability considerations and resource management, mostly in terms of clinical staff attention: A severe but stable patient might require less attention than a less severe patient that is quickly deteriorating. This presents an opportunity to design a prioritization of care prediction framework that takes into account other patients in the ICU. Such a framework could relieve the cognitive workload of clinicians caring for multiple patients simultaneously [33]. The output of prioritization algorithms could be integrated to existing systems such as the central EMR display, which displays patient vitals in

the nurses station. We note that since our study questions revolved around prediction systems addressing one patient at a time, further research is needed to fully understand the considerations involved in prediction for ICU prioritizing.

Our focus in this work was the algorithmic requirements from prediction models as they relate to the circumstances of its use. This analysis yielded at least two model schemes with two different aims (bedside model per patient, and an ICU-ward level model for prioritizing care). Since model integration into EHR systems and display considerations first depend on the final objective for its use, it was not studied in depth at this stage and should be the subject for further research.

5.2. Limitations

One limitation of this study is focusing on young adult patients with no comorbidities as the cases, and using a fixed narrow context for the quantitative tasks, while in reality clinicians would have had more information. Different contexts, such as age or comorbidities, could affect the results, as well as ongoing treatments. While we have touched on the issue of context in task *T3*, accounting for all of the above sources of variation would have required us to present to participants a much wider array of cases. This unfortunately was unrealistic given the time constraints of the clinicians who volunteered for the study. Moreover, the surprise aspect is only partially addressed by task *T3*, and it was not directly quantified in this study. Surprise also depends on context; it could be expressed as an unexpected sudden drop in a vital sign value, but also as a vital sign not improving despite a given treatment. This should be the subject of future work.

Additionally, trend was quantified in this study. While it was emphasized in the interviews and task *T1*, task *T4* reflected the variability of clinicians' views on the subject. We witnessed the entanglement of trend and absolute values in task *T1* (see Fig. 6). Since task *T4* displayed trajectories which do not exceed normal range, it is possible that clinicians could not separate the meaning of trend alone. Some clinicians specifically explained the low importance score they gave to some trajectories by referring to the value ending the trajectory.

Finally, sample size and diversity: we interviewed 13 people; while the themes that came up saturated after ~ 10 interview, the sample size makes it difficult to draw statistical conclusions, for example regarding the preferences of nurses vs. physicians, or preferences according to years of experience.

5.3. Conclusions

The intensive care environment is cognitively challenging, as clinicians are required to integrate vast amounts of information to assess and predict current and future patient states. For exactly these reasons, the ICU carries great potential for machine learning-based prediction models. Using a mixed-methods, interview-based study we explored the unique requirements of ICU clinicians from prediction systems in order to support a much needed user-centered design at the algorithmic level. Our findings highlight the types and properties of predictions clinicians value, the optimal timescales of such predictions, and uncover unmet needs regarding prioritization of care. Following these requirements, we propose ML practitioners spend more effort using prediction objectives which take into account patients' trajectory rather than events, allowing for less aggressive, preemptive interventions.

A recurring emerging theme, both from qualitative and quantitative evaluations, is that the perceived utility of models is context dependent, requiring integration of patient characteristics, short and medium term history, and ongoing clinical tasks. In a nutshell, the already cognitively-taxed clinicians value models that are actionable and create new knowledge or insights: "tell me something interesting that I don't know".

Finally, our results point the way to several directions for future research: The first is constructing evaluation metrics and loss functions for ICU prediction models which build on the derived utility functions. The second direction would be conducting another user study designed for display and interface planning for each of the new possible frameworks that came up in this study. Finally, we believe further research is needed to understand clinicians needs in care prioritization, moving beyond the single patient cases we used here and taking into account the entire patient population on in an ICU unit at a given time.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Azadeh Assadi for his insightful comments and suggestions which improved this work. We thank Sagi Porat for his assistance with tools used in this research, and Omer Eini for his contribution to the open-coding process. This research is funded by the Israeli Science Foundation and The Planning and Budgeting Committee of the Israeli Council for Higher Education.

Appendix A. Patient cases

In Fig. A.10 we present the ICU patient cases devised by our team, which includes an experienced ICU clinician. For each interview, we randomly selected 2–3 cases below to lead the interview. These are standard ICU scenarios with varying degrees of severity.

Appendix B. T1: Identifying important events - extended

Additional examples for task *T1* displayed in Fig. B.11. While sensitivity changed among participants, the same three considerations were reported for marking significant events: deviation from clinical norm, overall trend, and surprising trend deviations. (See Figs. B.12,B.13 and B.14).

Additionally, we present examples from of *T1*'s raw data corresponding to results presented at Section 3.2.2:

- Clinical severity is determined by the absolute values and the trend of the presented vital sign *["These values requires ventilation, this is severe" -N1], ["This trend is important because if it goes any lower then something bad will happen to this patient" -P3*]
- High or low absolute values are important as they describe clinically severe states - *"This is very important.. With these readings.. this is clinically severe" -N1*
- General trend hints at clinical severity even if the absolute values are not severe just yet – *"This downward trend is bad.. it is also long, may the staff is missing somethong" -N1*.

- As for sudden deviations from the trend, participants provided two different explanations for their importance: One is that a sudden change could be a symptom of a clinical event *"Such sudden change describes an event, something snapped" -P6*. The other is that these are less likely to be predicted by the staff and therefore they are a prediction of a surprising event *"Surprise is anything that deviates from my expectations" -N4*.
- The same slope of heart rate trajectory is more important when the absolute values are also concerning: *"This whole trend is important, but after it goes down below 92 it is very important to me" -P6*

Appendix C. T3: Anchoring

As mentioned in *Methods*, the two versions of this task achieved the same result. A change in the importance curve occurs only if previous measurements exceed clinical norm and are considered cause for concern (at least in the terms revealed in task *T2*). Fig. C.15 displays measurements that are borderline according to the medical literature but were not rated as concerning in task *T2* and yield no change with respect to the baseline curves.

Appendix D. T4: Pure trend

The curves in Fig. D.16 represent the importance attributed by the participants to displayed trends in task *T4*. Importance seems to increase non-linearly as vitals' slope becomes steeper. We note the SpO2 curve presents differently since it has a single direction of exceeding norm – only towards lower values.

Appendix E. Coding table

Prediction Content
Vitals Trajectory
Predict trend
Bad trend
State Trajectory
Changes
Window
Step predictions
Multiple future states scores
Multiple future vitals readings summaries
Counterfactual predictions
Events
Critical
Minor
Sudden vital change
Requires non-aggressive treatment
Trend Deviation
Prioritizing Care
Respond
Monitor
Time
High frequency
monitoring
Predictions
Two prediction timelines
Days
Minutes
Timeline per disease
Disease time
The shortest time step of existing diseases
Context
Surprise
Abnormal attributes
Unexpected trajectory
Patient baseline
Probability
Pros
Cons
Performance evaluation per patient
Consider extensive context (meds etc.)
Clinically severe (bad trend, not recovering etc.)
Workflows
Concern for false alarms
Frequent re-evaluations
Accounting for "soft indicators"
Nighttime overload
Human errors
Reluctant to maintain information systems
Other Machine Learning Opportunities
Diagnostics
Action recommendations
Roles & Experience

(*continued on next page*)

(*continued*)

Physician
Nurse
Intern
Senior
Adults
Pediatrics

References

[1] S. Blecker, D. Sontag, L.I. Horwitz, G. Kuperman, H. Park, A. Reyentovich, S. D. Katz, Early Identification of Patients With Acute Decompensated Heart Failure, J. Cardiac Fail. 24 (2018) 357–362.

[2] A.E.W. Johnson, M.M. Ghassemi, S. Nemati, K.E. Niehaus, D. Clifton, G.D. Clifford, Machine Learning and Decision Support in Critical Care, Proc. IEEE 104 (2016) 444–466.

[3] K.E. Henry, D.N. Hager, P.J. Pronovost, S. Saria, A targeted real-time early warning score (TREWScore) for septic shock, Sci. Translat. Med. 7 (2015) 122–299.

[4] M. Schvetz, L. Fuchs, V. Novack, R. Moskovitch, Outcomes prediction in longitudinal data: Study designs evaluation, use case in ICU acquired sepsis, J. Biomed. Inform. 117 (2021) 103734.

[5] Y. Si, J. Du, Z. Li, X. Jiang, T. Miller, F. Wang, W.J. Zheng, K. Roberts, Deep representation learning of patient data from electronic health records (EHR): A systematic review, J. Biomed. Inform. 115 (2021) 103671.

[6] M. Moor, B. Rieck, M. Horn, C.R. Jutzeler, K. Borgwardt, Early prediction of sepsis in the ICU using machine learning: a systematic review, Front. Med. 8 (2021) 348.

[7] T. Desautels, J. Calvert, J. Hoffman, M. Jay, Y. Kerem, L. Shieh, D. Shimabukuro, U. Chettipally, M.D. Feldman, C. Barton, D.J. Wales, R. Das, Prediction of Sepsis in the Intensive Care Unit With Minimal Electronic Health Record Data: A Machine Learning Approach, JMIR Med. Informat. 4 (2016) e28.

[8] S. Ghosh, J. Li, L. Cao, K. Ramamohanarao, Septic shock prediction for ICU patients via coupled hmm walking on sequential contrast patterns, J. Biomed. Inform. 66 (2017) 19–31.

[9] A.-S. Poncette, C. Spies, L. Mosch, M. Schieler, S. Weber-Carstens, H. Krampe, F. Balzer, Clinical requirements of future patient monitoring in the intensive care unit: qualitative study, JMIR Med. Informat. 7 (2019) e13064.

[10] O. Faust, Y. Hagiwara, T.J. Hong, O.S. Lih, U.R. Acharya, Deep learning for healthcare applications based on physiological signals: A review, Comput. Methods Programs Biomed. 161 (2018) 1–13.

[11] L. Clifton, D.A. Clifton, M.A. Pimentel, P.J. Watkinson, L. Tarassenko, Gaussian processes for personalized e-health monitoring with wearable sensors, IEEE Trans. Biomed. Eng. 60 (2012) 193–197.

[12] P. Schulam, S. Saria, A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure, in: Advances in Neural Information Processing Systems, vol. 28.

[13] A.M. Alaa, J. Yoon, S. Hu, M. Van der Schaar, Personalized risk scoring for critical care prognosis using mixtures of Gaussian processes, IEEE Trans. Biomed. Eng. 65 (2017) 207–218.

[14] H. Soleimani, A. Subbaswamy, S. Saria, Treatment-response models for counterfactual reasoning with continuous-time, continuous-valued interventions, in: Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence, 2017.

[15] G.W. Colopy, S.J. Roberts, D.A. Clifton, Bayesian optimization of personalized models for patient vital-sign monitoring, IEEE J. Biomed. Health Informat. 22 (2017) 301–310.

[16] L.-F. Cheng, B. Dumitrascu, G. Darnell, C. Chivers, M. Draugelis, K. Li, B. E. Engelhardt, Sparse multi-output Gaussian processes for online medical time series prediction, BMC Med. Informat. Decision Making 20 (2020) 1–23.

[17] N.L. Downing, J. Rolnick, S.F. Poole, E. Hall, A.J. Wessels, P. Heidenreich, L. Shieh, Electronic health record-based clinical decision support alert for severe sepsis: a randomised evaluation, BMJ Quality Saf. 28 (2019) 762–768.

[18] A. Wong, E. Otles, J.P. Donnelly, A. Krumm, J. McCullough, O. DeTroyer-Cooley, J. Pestrue, M. Phillips, J. Konye, C. Penoza, et al., External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients, JAMA Int. Med. 181 (2021) 1065–1070.

[19] J.L. Guidi, K. Clark, M.T. Upton, H. Faust, C.A. Umscheid, M.B. Lane-Fall, M. E. Mikkelsen, W.D. Schweickert, C.A. Vanzandbergen, J. Betesh, et al., Clinician perception of the effectiveness of an automated early warning and response system for sepsis in an academic medical center, Ann. Am. Thoracic Soc. 12 (2015) 1514–1519.

[20] S. Tonekaboni, S. Joshi, M. D. McCradden, A. Goldenberg, What clinicians want: Contextualizing explainable machine learning for clinical end use, in: Proceedings of the 4th Machine Learning for Healthcare Conference, Proceedings of Machine Learning Research, PMLR, vol. 106, 2019, pp. 359–380.

[21] O. Amir, B. J. Grosz, K. Z. Gajos, S. M. Swenson, L. M. Sanders, 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, Association for Computing Machinery, New York, NY, USA, 2015, pp. 1419–1428.

[22] M. Jacobs, J. He, M. F. Pradier, B. Lam, A. C. Ahn, T. H. McCoy, R. H. Perlis, F. Doshi-Velez, K. Z. Gajos, Designing ai for trust and collaboration in time-constrained medical decisions: A sociotechnical lens, in: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–14.

[23] G.J. Escobar, B.J. Turk, A. Ragins, J. Ha, B. Hoberman, S.M. LeVine, M.A. Ballesca, V. Liu, P. Kipnis, Piloting electronic medical record–based early detection of inpatient deterioration in community hospitals, J. Hospital Med. 11 (2016) S18–S24.

[24] M. C. Elish, The stakes of uncertainty: developing and integrating machine learning in clinical care, in: Ethnographic Praxis in Industry Conference Proceedings, vol. 1, Wiley Online Library, pp. 364–380.

[25] S. Malhotra, D. Jordan, E. Shortliffe, V.L. Patel, Workflow modeling in critical care: piecing together your own puzzle, J. Biomed. Inform. 40 (2007) 81–92.

[26] V.L. Patel, J. Zhang, N.A. Yoskowitz, R. Green, O.R. Sayan, Translational cognition for decision support in critical care environments: a review, J. Biomed. Inform. 41 (2008) 413–431.

[27] A.-S. Poncette, L. Mosch, C. Spies, M. Schmieding, F. Schiefenhövel, H. Krampe, F. Balzer, et al., Improvements in patient monitoring in the intensive care unit: survey study, J. Med. Internet Res. 22 (2020) e19091.

[28] T. Foster-Hunt, A. Parush, J. Ellis, M. Thomas, J. Rashotte, Information structure and organisation in change of shift reports: An observational study of nursing hand-offs in a paediatric intensive care unit, Intensive Crit. Care Nurs. 31 (2015) 155–164.

[29] R. Jääskeläinen, Think-aloud protocol, Handbook Translat. Stud. 1 (2010) 371–374.

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

[31] A. Strauss, J. Corbin, Basics of qualitative research, Sage Publications (1990).

[32] K. Holtzblatt, Contextual design, in: The human-computer interaction handbook, CRC Press, 2007, pp. 975–990.

[33] K. Jung, S. Kashyap, A. Avati, S. Harman, H. Shaw, R. Li, M. Smith, K. Shum, J. Javitz, Y. Vetteth, et al., A framework for making predictive models useful in practice, J. Am. Med. Inform. Assoc. 28 (2021) 1149–1158.

[34] H. Suresh, N. Hunt, A. Johnson, L. A. Celi, P. Szolovits, M. Ghassemi, Clinical intervention prediction and understanding with deep neural networks, in: F. Doshi-Velez, J. Fackler, D. Kale, R. Ranganath, B. Wallace, J. Wiens (Eds.), Proceedings of the 2nd Machine Learning for Healthcare Conference, Proceedings of Machine Learning Research, PMLR, vol. 68, 2017, pp. 322–337.

[35] A.M. Alaa, M. van der Schaar, Bayesian inference of individualized treatment effects using multi-task Gaussian processes, Adv. Neural Inform. Process. Syst. 30 (2017).

[36] A. Faiola, C. Newlon, Advancing critical care in the ICU: a human-centered biomedical data visualization systems, in: International Conference on Ergonomics and Health Aspects of Work with Computers, Springer, Berlin, Heidelberg, pp. 119–128.

[37] A. Faiola, P. Srinivas, J. Duke, Supporting clinical cognition: a human-centered approach to a novel ICU information visualization dashboard, in: AMIA Annual Symposium Proceedings, vol. 2015, American Medical Informatics Association, p. 560.

[38] J.R. Busemeyer, J.T. Townsend, Decision field theory: a dynamic-cognitive approach to decision making in an uncertain environment, Psychol. Rev. 100 (1993) 432.

[39] S. Yang, K. Kalpakis, C. F. Mackenzie, L. G. Stansbury, D. M. Stein, T. M. Scalea, P. F. Hu, Online recovery of missing values in vital signs data streams using low-rank matrix completion, in: 2012 11th International Conference on Machine Learning and Applications, vol. 1, IEEE, pp. 281–287.

[40] O. Linial, N. Ravid, D. Eytan, U. Shalit, Generative ODE modeling with known unknowns, in: Proceedings of the Conference on Health, Inference, and Learning, pp. 79–94.

[41] Pragasan Dean Gopalan, Santosh Pershad, Decision-making in ICU – A systematic review of factors considered important by ICU clinician decision makers with regard to ICU triage decisions, Journal of Critical Care (2019). In press.

[42] Q. Yang, A. Steinfeld, J. Zimmerman, Unremarkable ai: Fitting intelli-650gent decision support into critical, clinical decision-making processes, in:Proceedings of the 2019 CHI Conference on Human Factors in ComputingSystems, CHI '19, Association for Computing Machinery, New York, NY,USA, 2019, p. 1–11.

[43] A. Wassenaar, L. Schoonhoven, J.W. Devlin, van Haren, A.J. Slooter, P.G. Jorens, M. & van den Boogaard, Delirium prediction in the intensive care unit: comparison of two delirium prediction models, Critical Care 22 (1) (2018) 1–9.