Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care

Venkatesh Sivaraman
venkats@cmu.edu
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA

Leigh A. Bukowski
University of Pittsburgh, School of Medicine
Pittsburgh, Pennsylvania, USA

Joel Levin
University of Pittsburgh
Pittsburgh, Pennsylvania, USA
joel.levin@pitt.edu

Jeremy M. Kahn
University of Pittsburgh and UPMC Health System
Pittsburgh, Pennsylvania, USA
jeremykahn@pitt.edu

Adam Perer
Carnegie Mellon University
Pittsburgh, Pennsylvania, USA
adamperer@cmu.edu

ABSTRACT

Artificial intelligence (AI) in healthcare has the potential to improve patient outcomes, but clinician acceptance remains a critical barrier. We developed a novel decision support interface that provides interpretable treatment recommendations for sepsis, a life-threatening condition in which decisional uncertainty is common, treatment practices vary widely, and poor outcomes can occur even with optimal decisions. This system formed the basis of a mixed-methods study in which 24 intensive care clinicians made AI-assisted decisions on real patient cases. We found that explanations generally increased confidence in the AI, but concordance with specific recommendations varied beyond the binary acceptance or rejection described in prior work. Although clinicians sometimes ignored or trusted the AI, they also often prioritized aspects of the recommendations to follow, reject, or delay in a process we term "negotiation." These results reveal novel barriers to adoption of treatment-focused AI tools and suggest ways to better support differing clinician perspectives.

CCS CONCEPTS

• **Human-centered computing** → **Interactive systems and tools**; • **Applied computing** → **Health informatics**.

KEYWORDS

human-AI interaction, healthcare, visualization, interpretability

ACM Reference Format:
Venkatesh Sivaraman, Leigh A. Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany*. ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3544548.3581075

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04.
https://doi.org/10.1145/3544548.3581075

1 INTRODUCTION

Artificial intelligence (AI) in health care promises to improve outcomes, reduce costs, and save clinicians time and effort. Yet at present, even the most encouraging AI solutions face significant obstacles to deployment and acceptance in real-world clinical settings [59]. AI-based tools that seek to improve decision-making across diverse deployment contexts must produce recommendations that are both acceptable to health care providers and transparent in the case of errors [30, 82]. In addition, health care providers generally consider themselves to be content experts in their fields, and they are naturally skeptical of decision aids that may limit their autonomy and challenge their sense of identity [70]. These issues have motivated studies that aim to evaluate and improve the human-AI collaborative system in health care [35, 40, 53], often drawing on insights from interpretable AI [52, 67, 85]. By helping clinical experts understand the conditions in which AI predictions fail, clinical decision support (CDS) systems could help produce AI-assisted decisions that are better than those made by humans or algorithms alone, improving patient outcomes.

Despite these efforts, effective complementarity between humans and AI-based CDS has largely not yet been realized, in part because it is difficult for clinicians to calibrate their trust in newly developed AI systems. Experimental studies demonstrate that trust can be miscalibrated in both directions: experienced clinicians often dismiss AI recommendations regardless of quality, while novices over-rely on incorrect advice [7, 31]. Prior work has investigated several strategies to mitigate these discrepancies, including providing explanations of the process underlying a given recommendation [1, 12, 85], communicating the uncertainty of predictions [78, 90], and familiarizing users with the AI's global strengths and weaknesses as identified from external validation [14, 32]. However, none of these methods appear to work universally across contexts, particularly in health care; in some domains, they may inflate or undermine confidence [31, 40, 90], while in others they may be poorly-aligned with human decision-making processes [2, 32]. In order to cultivate an appropriate level of reliance, interpretable AI systems must account for decision-maker characteristics, task complexity, AI performance, and other factors in ways that are not yet fully understood [52, 57, 67].

Importantly, most work in this area has focused on *diagnostic* systems—that is, systems designed to help clinicians make a clinical

diagnosis (e.g., identifying cancer in a radiograph [16]) or predict a future clinical event (e.g., clinical deterioration [35, 68]). These systems have the benefit of a known or expert-annotated "ground truth" upon which algorithms can be developed and calibrated. As such, they aim to improve clinical care by reducing diagnostic errors, providing insight into the likelihood of future events, and minimizing cognitive burden.

A separate, emerging class of CDS systems is designed to deliver *treatment* recommendations—for example, which type of chemotherapy to give to a cancer patient or whether or not to administer intravenous fluids to a hospitalized patient identified as being at-risk. In these systems, the "best" decisions are often difficult to identify, either due to a lack of clinical evidence [22, 86] or due to expert disagreement on the best course of action [24, 63]. Therefore, treatment-focused AI models often seek to discover optimal decisions by correlating treatments with their average effects on patient outcomes, a task which is significantly more challenging than diagnosis or prediction but which has the potential to more powerfully impact patients.

While clearly different from a modeling perspective, a key question is whether AI tools that do not have a clear correct decision require different approaches for *human-AI interaction design*. For diagnostic aids and risk-assessment tools, AI-assisted decision-making behavior is often conceptualized as taking place at a single point in time (e.g. a physician encounter) and as involving a limited number of options (e.g. either agreement or disagreement with a diagnosis) [14, 16, 35, 45, 80, 83]. In contrast, decisions about *treatments* often span a wider range of options and take place across multiple time-points that can confound outcomes. For instance, a clinician may prescribe one treatment on the first visit, then observe that it is having little effect on the patient's condition and administer a different treatment upon the second visit. Determining the optimal decision in this context requires understanding both how past treatment decisions have affected the current patient state, and how future treatment decisions might influence the expected outcome. This complex reasoning task is well-known in causal modeling [60] and has been described in early-stage design studies in health care [46]. It has not, to our knowledge, been explored in the context of a functioning treatment decision support tool.

In this work, we sought to explore how clinicians interact with real AI-based treatment recommendations in a setting where sequential treatments can affect outcomes in complex ways [46]. Our clinical domain of interest was the intensive care unit (ICU), an environment characterized by acutely ill hospitalized patients and correspondingly dynamic, time-sensitive decisions. We designed and implemented an interactive CDS interface that delivers interpretable recommendations for treating sepsis, a life-threatening medical condition with relatively few existing evidence-based care protocols and substantial heterogeneity in treatment patterns among clinicians [17]. The foundation of our CDS was an existing well-known AI model that could reduce patient mortality if followed [50] but that has not been prospectively evaluated in a clinical setting. The resulting CDS system formed the basis for a think-aloud study with 24 clinicians, all of whom practice in the ICU and have experience treating sepsis. We aimed to understand their responses to the recommendations and explanatory evidence provided by the AI, including both how they perceived it to influence their decisions and how their actual treatment choices were affected.

A mixed-methods analysis of the think-aloud transcripts and structured decision responses showed that explanations improved clinicians' perceptions of the AI's usefulness and made them more confident in their own decisions, a finding that is consistent with prior literature [1, 80, 85]. However, their overall rates of binary concordance with the AI recommendations did not appear to be affected by explanatory visualizations. Instead, analysis of participants' think-aloud decision processes revealed a more nuanced picture of individual decision-making than described in the literature thus far, involving four distinct behavior patterns with the AI:

(1) **Ignore,** in which the decision-maker is not affected by the AI recommendation in any decision;
(2) **Negotiate,** in which the decision-maker weighs and prioritizes individual aspects of the recommendation to follow or adjust;
(3) **Consider,** in which the decision-maker dichotomously defers to or overrides the recommendation; and
(4) **Rely,** in which the decision-maker accepts some part of the recommendation in every decision.

These behavior patterns, particularly in the Negotiate group, indicate that recommendations for treatment decisions in the ICU may be subject to *partial* forms of reliance that could impact the efficacy of chosen treatments in undetermined ways. Our results also pointed to ways in which the formulation of the model used for this study hindered clinicians from using it effectively, opening new directions for model improvement and evaluation. We discuss the implications of the behavior patterns and obstacles we observed on the further development of AI-based treatment decision support tools.

2 BACKGROUND AND RELATED WORK

2.1 Sepsis Diagnosis and Treatment in the ICU

Sepsis is a life-threatening medical condition that affects over 1.7 million adults in the United States each year and is the leading cause of death in hospitals [18]. Sepsis occurs when the body's response to an infection causes systemic inflammation and organ dysfunction [28], which can in turn lead to septic shock and death [71]. Sepsis is also the most costly condition in U.S. hospitals [10], and as such it represents a major target for quality improvement efforts at the local and national level [36]. Timely identification and appropriate management of sepsis is crucial to reducing mortality rates [28]. Key diagnostic strategies include frequent clinical assessments, blood cultures to identify pathogens, and measurement of laboratory values that may indicate infection; treatment strategies include control of the infectious source with antibiotics or antivirals, intravenous (IV) fluids to maintain appropriate fluid balance, and vasopressors (such as norepinephrine) to maintain appropriate blood pressure [28].

Sepsis has long been a focus of AI research; however, nearly all of this research is devoted to early identification and diagnosis. Multiple machine learning algorithms exist for mining hospital electronic health record data to identify patients with sepsis [58, 64, 77]. These systems are generally accurate, and several hospital systems

have already implemented algorithmic early warning systems for sepsis [35, 69]. However, the implementation of these systems has not tended to affect treatment decisions or patient outcomes [38]. Conceptually, early warning systems may fall short of their goal of improving the quality of care when they fail to provide information that is both novel and actionable.

In contrast, relatively little AI research has focused on sepsis treatment. Guidelines for treating sepsis in the ICU are continually evolving [28], and although individual treatment decisions at specific time points (e.g., whether to give fluids or vasopressors) are certainly highly influential on patient outcomes *on average*, the physiological complexity of sepsis renders the influence of treatments on individual outcomes largely unknowable. As such, recommendations face significant challenges in translation to wider clinical practice [75], resulting in substantial variability in care practices [6] and continued high mortality levels [74].

Machine learning approaches for sepsis treatment aim to standardize and improve sepsis care by leveraging historical patient trajectories. The most prominent example of this approach is the AI Clinician developed by Komorowski et al. [50], and it is the model that forms the basis for our clinician-facing work. By improving the consistency and timeliness around treatment with IV fluids and vasopressors, sepsis treatment models such as the AI Clinician have immense potential to reduce mortality (from around 13% to around 5%, according to [29]). However, for this potential to be realized, clinicians actually have to act on the AI recommendations at the bedside. Indeed, studies evaluating the effects of these predictions have only considered retrospective data and not how (or if) such tools might be utilized by human clinicians. Because model recommendations are impactful only if they are implemented, it is critical to understand how a model like the AI Clinician might be integrated into an ICU clinician's workflow, and whether human-AI collaboration can indeed outperform unaided human clinicians.

2.2 Explainability, Interpretability and Decision-Making

The design of explainable and interpretable ML-based tools has become a major focus of research in the HCI community [26, 76, 90]. While early efforts in explainable AI (XAI) focused on feature-based explanations, current conceptions of interpretability comprise a wider range of techniques, including uncertainty and confidence metrics [65], nearest-neighbors [39], and counterfactuals [91]. In concert with human-centered design methods, these technical approaches can be integrated into algorithmic systems with the intent of improving trust and human-AI team performance [5, 37, 67].

However, there remain substantial challenges in designing and validating interpretable AI systems, particularly in high-stakes decision-making domains such as health care. Model explanations themselves can be prone to issues such as over-sensitivity to input values or giving seemingly-sensible explanations for incorrect predictions [73]. When explanations are presented to decision-makers alongside predictions, a line of studies ranging from Bussone et al. [12] to more recent work [20, 85, 90] has shown that these explanations tend to increase trust in the model even when it is unwarranted. Explanations can also interact with reasoning fallacies such as confirmation and availability bias [21, 45], but mitigating these effects requires knowledge of a normatively correct reasoning process [84] that may not always be available.

In the translation of promising AI tools into a real-world setting, the evaluation of decision quality poses its own challenges. While expert consensus can serve as a useful proxy for the ground truth [40], the accuracy of a real-world decision is often fundamentally unknowable and contentious [47]. Taking an alternative strategy, some AI systems instead strive to provide clinicians with useful non-prescriptive information, such as highlighting informative parts of a medical image [27], displaying information from similar historical cases [13], or identifying patients at high risk of future deterioriation [35, 68]. These approaches serve to focus attention without making specific recommendations, thereby indirectly improving decisions but also making the AI algorithms more ignorable (potentially reducing benefit).

The present study was specifically designed to address the challenges described above: imperfect and biased models, potentially misleading and hard-to-interpret explanations, and in particular the lack of an objective ground truth. Rather than evaluating decision-making along a single axis of quality, we used a combination of behavioral and attitudinal measures [66] to understand how an imperfect AI system would affect its users within a high-stakes environment in which the correct decision is unknowable in real-time.

2.3 Clinician Perceptions of Decision Support Tools

While ML-based tools for clinical decision-making have great potential utility, they face the combined challenges of building effective XAI as well as broader obstacles to adoption of CDS tools. Yang et al. [87] describe difficulties in gaining acceptance from expert clinicians without formal validation of the tool, as well as the inherent challenge of situating CDS at the right time and place for decision-making. Similarly, Cai et al. [14] emphasize clinicians' need to understand the overall design and validation of the CDS before they can trust it on individual instances. Studies of deployed AI systems by Beede et al. [9] and Wang et al. [83] have identified clinician frustrations with the added workload of using a CDS, particularly when those systems do not adequately complement their expertise. Early-stage studies of CDS tools have also found that acceptance of AI recommendations is often more strongly determined by the clinician's expertise than by the quality of the recommendation [31, 80].

On the other hand, a few deployed systems have met with success and clinician acceptance. For example, AI-driven CDS systems for image-based diagnosis have been increasingly accepted as tools to reduce clinician burden and prioritize attention [9, 80]. Related to sepsis, early-warning systems such as Sepsis Watch and the Targeted Real-time Early Warning System (TREWS) have been accepted by clinicians at the hospitals where they are deployed [35, 69], despite being initially met with ambivalence [34]. These tools may have been readily accepted because (1) they could be rigorously validated using ground-truth data, (2) they ultimately helped coordinate providers and prioritize care [35] rather than directly replacing clinical judgment; and (3) acceptance does not rely on clinician behavior change and they can therefore be easily ignored by

untrusting clinicians. However, while these diagnosis-focused tools are a promising model for AI-based CDS, they represent only one of many points in the care workflow in which complex decisions may be needed.

In particular, relatively few studies have examined the acceptability of AI-generated *treatment recommendations*: Jacobs et al. used a mock AI system to evaluate clinician decisions on antidepressant selection [40], while Yang et al. presented clinicians with projected survival curves conditioned on a device implantation decision [87]. Kaltenhauser et al. examined intravenous fluid administration in intensive care, although their study was more focused on understanding decision-making without an AI rather than the influence of AI on treatment decisions [46]. In contrast to diagnoses, which are relatively straightforward to learn from historical data, treatment recommendations in CDS have predominantly been derived from clinical best-practice guidelines rather than AI [8, 43]. However, broad best-practice guidelines are of limited utility at the bedside because their recommendations fail to account for patient-level variation and the interaction between multiple variables over time [49]. Machine learning approaches have the potential to deliver treatment recommendations that are more specific and personalized, but how they will be received by clinicians remains an open question.

3 DESIGN OF AN INTERACTIVE AI-DRIVEN CDS SYSTEM

As an initial step towards bringing AI-based recommendations to clinical practice for sepsis treatment, we designed and implemented an interactive patient trajectory visualization tool called the AI Clinician Explorer. This tool serves as both an exploratory tool for historical patient data and as an interface for a real treatment recommendation model developed from retrospective clinical data (the AI Clinician). The following sections describe the underlying model as well as the design of the front-end visualization system.

3.1 Reinforcement Learning for Sepsis Treatment

Unlike many AI problem formulations which treat the patient as a static data point on which to make a prediction, sepsis management in the ICU requires a dynamic approach that considers the changing state of the patient over time. In particular, models need to account for the fact that long-term outcomes, such as mortality, may not be the direct result of a single action but rather a series of actions over an evolving trajectory. The AI Clinician [50] addresses these challenges by applying a reinforcement learning (RL) strategy. Like most RL approaches in health care, the AI Clinician works by modeling a patient trajectory as a sequence of memoryless states derived from available biometric signals (vital signs, lab values, etc.). At each timestep, the agent—either a clinician or a model—can choose from a predefined set of actions, which then results in the agent receiving a numerical reward (or penalty). As shown in Fig. 1, the AI Clinician uses k-means clustering to define 750 possible patient states, then applies an algorithm called policy iteration to determine which of 25 different treatment actions most optimally reduces mortality in each state. The model's actions represent 5 possible levels of IV fluids and 5 vasopressor dosages binned by quantiles, representing a substantial but non-exhaustive subset of the treatment choices that a human clinician might make.

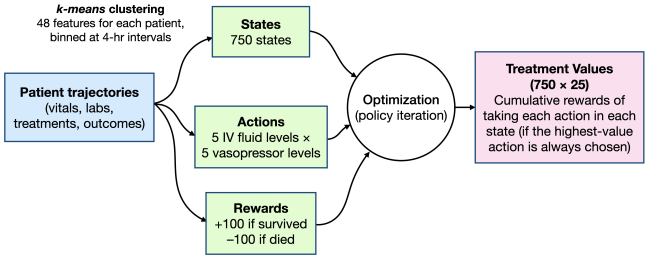

Figure 1: Overview of the AI Clinician's training methodology, summarized from [50]. The model takes as input a set of historical trajectories comprising patient vitals, labs, and treatments discretized at 4-hour intervals. Each timestep is represented as one of 750 different states (determined using clustering) followed by one of 25 possible treatment actions. The output of the model is a set of treatment values (or Q-values), which estimate the future rewards that would be obtained from taking a given action. The *policy* that the AI Clinician would follow is to take the action with the largest value estimate in each state.

The AI Clinician publication is widely known and highly influential in both the critical care and CDS communities [79, 88]. Yet it has also faced criticism because its recommendations often deviate from bedside clinicians' best understanding. This may be due to inherent biases in how patients' outcomes are weighted in the model's evaluation process, a problem known in RL as off-policy evaluation [41]. Additionally, more recently-developed techniques using deep neural networks may improve on the accuracy of the AI Clinician [29, 61, 89]. However, since these more recent methods also rely on off-policy evaluation, reliable benchmarks of their performance remain elusive. For this study, we chose Komorowski et al.'s approach because it is the best-known model of its kind, and therefore most likely to gain clinician acceptance in the absence of concrete evidence that any such AI model improves outcomes.

We replicated the AI Clinician's methodology using the publicly-available MIMIC-IV dataset [42] (a more recent version of the MIMIC-III dataset used by the original model developers), and provide the code on GitHub for future reproducibility[1]. MIMIC includes granular clinical data on all ICU admissions to a large academic medical center over a multi-year time period, and therefore is a unique resource for this project. The model was trained on a cohort of 18,143 patients who met standard diagnostic criteria for sepsis at some point during their ICU stay. We verified that the model performance on held-out data was similar to the original reported performance, as measured by a bootstrapped policy value estimate computed using weighted importance sampling[2]. Specifically, the model whose predictions were displayed had a policy value of 83.8

[1] https://github.com/cmudig/AI-Clinician-MIMICIV

[2] The accuracy of an RL policy cannot be computed directly on retrospective data because we cannot observe the outcomes of following the policy. Instead, weighted importance sampling (WIS) works by averaging the survival/mortality rewards associated with each trajectory, weighted by how similar the clinicians' actions were to the model predictions.

(possible values range from −100 to 100), while the values reported by Komorowski et al. on MIMIC-III ranged between 80 and 90 [50].

3.2 Visualization System Design

We next developed a novel front-end visualization system which we term the AI Clinician Explorer. This system enables clinical experts to search for patients in the MIMIC-IV dataset, visualize their disease trajectories, and compare model predictions to actual treatment decisions delivered at the bedside. The AI Clinician Explorer was designed for use both as a tool for research and education on AI in sepsis, and as a starting point for an eventual clinician-facing interface for real-time decision-making in the live clinical environment. An initial prototype was created using inspiration from prior literature, notably ClinicalVis [33]. We then iterated on this design and tailored it for use by ICU clinicians, based on feedback from experienced ICU physicians and other experts in biomedical sciences, informatics, and psychology. The final system consists of the following primary components:

Browse and filter patients. The Patient Browser page helps users find cases of interest by allowing them to filter and sort a list of patient trajectories by a variety of task-specific metrics. The most straightforward of these include filters for age, gender, comorbidities, outcomes, and commonly-used disease severity scores (SOFA and SIRS). During the iterative development process, we identified a need to filter for specific actions and recommendations at a timestep level (i.e., the 4-hour time periods over which the model aggregates data and makes treatment recommendations), which is more granular than filtering at the patient-level. We added filter controls that allow the user to select from the 25 possible clinician and model actions on a pair of grids. This was used to identify timesteps in which, for example, clinicians tended to give IV fluids while the model recommended vasopressors.

Visualize patient trajectory. The Patient Trajectory page, represented in Fig. 2, was designed to help clinicians quickly assess a patient's state throughout their ICU stay. Similar to ClinicalVis [33], our trajectory visualizations communicate the patient's current vital signs and lab values numerically, and depict their trends over time using line charts. Abnormal values are highlighted in red, while trend arrows show changes in each value relative to the last 4 hours. Clinicians' feedback on the time-series charts indicated that the visualizations were highly usable, particularly compared to how data is currently presented in existing hospital-based electronic health records (EHRs). We also worked with the clinicians to reorder and regroup the features into semantic categories, which served to align the page's structure with standard reporting conventions and facilitate skimming.

Compare model predictions and clinician actions. As described in Sec. 3.1, the AI Clinician categorizes each patient to one of 750 states, and each state is associated with a predicted treatment recommendation. For each timestep in each patient trajectory, the interface displays heatmaps showing the AI Clinician's recommended action (Fig. 2c) and the distribution of historical clinician actions in that state (Fig. 2d). This pair of visualizations surfaced the insight that the AI Clinician often assigns similar treatment values to multiple actions rather than strongly preferring a single action. We therefore explored ways to present multiple treatment options during the study, resulting in the Alternative Treatments visualization condition.

Interpret state clustering. Presenting explanations of the AI Clinician's predictions was a key aspect of both the patient browser interface and the clinician-facing study. A few explainability methods have been developed specifically for reinforcement learning (XRL) [55, 62]; however, these methods generally either require different training methods (to create intrinsically interpretable policies) or accurate models of patient trajectory dynamics (to predict counterfactual outcomes). To align our work with both the existing AI Clinician model and prior XAI literature [85, 90], we opted to use standard XAI techniques to explain the state clustering, which has a major impact on the model output as it determines which groups of patients are recommended similar treatments. For each of the 750 possible states, we trained an XGBoost classifier [19] to predict whether a patient was in that state or not. We then used Shapley Additive Explanations (SHAP) [54] to identify the features that most often contributed to patients being included in the state.[3] These features are depicted in the State Interpretation chart, and were used in the Feature Explanation study condition.

We used the AI Clinician Explorer to select patients for our clinician-facing study, as described in greater detail in Sec. 4.1. The system also formed the basis of the interface that study participants used to make decisions. (In the study we controlled which model visualizations and time points participants saw rather than giving them access to the entire AI Clinician Explorer, thus better replicating the information set that would be available to clinicians in real life.) The tool is built using a Flask back-end, a Svelte front-end, and a database comprising BigQuery and Google Cloud Firestore components. The source code for the tool and study interface is available on GitHub to support future research[4].

4 STUDY METHODS

We conducted a mixed-methods study to understand the challenges that clinicians face when attempting to incorporate AI advice, to explore how participants perceive their decision-making differently with AI support, and to evaluate the effect of explanations on acceptance. We explored the following research questions:

(1) How do AI-generated sepsis treatment recommendations and explanatory visualizations affect clinicians' perceptions of decision-making?
(2) How do visual explanations of model predictions affect acceptance of the AI's advice?
(3) What challenges do clinicians perceive in incorporating AI treatment recommendations into their decision-making?

We recruited 24 practicing ICU clinicians from a large multi-hospital academic hospital system in the eastern United States. Our sample included three types of ICU clinicians representing the range of providers that make sepsis treatment decisions in the ICU: attending physicians, advanced practice providers (APPs), and critical care fellows in training. Attending physicians are the most

[3]The choice of explanation technique can have a significant effect on what conclusions are drawn from feature importance charts. In our case, the combination of XGBoost and SHAP qualitatively yielded more parsimonious, clinically sensible explanations than other classifiers (random forests, SVMs) or explanation techniques (SVM coefficients, permutation importance).
[4]https://github.com/cmudig/ai-clinician-explorer

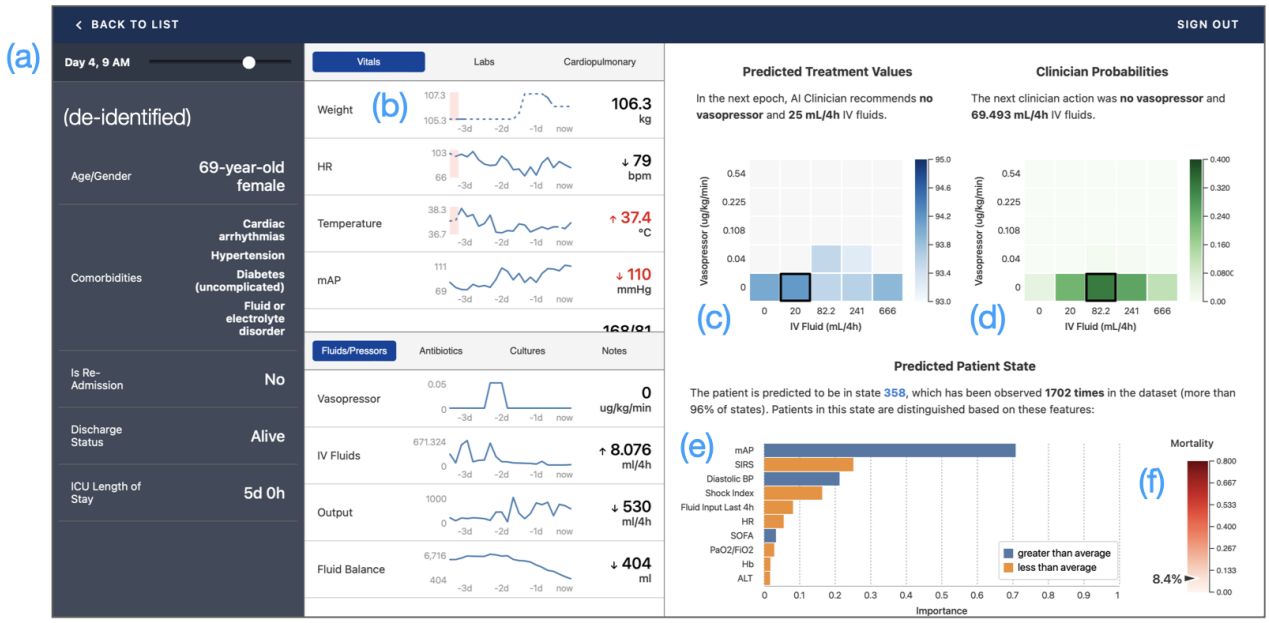

Figure 2: Main interface in the AI Clinician Explorer, designed to support browsing patient trajectories, interpreting model recommendations and comparing predicted treatment values against historical clinical actions. (a) The timestep control allows the user to step through the patient's ICU stay at 4-hour intervals. (b) The patient state is shown in small-multiple line charts, with abnormal values highlighted in red. (c) Heatmap showing the estimated value of taking each of the 25 possible actions (Q values) from the current state. (Values are only estimated for actions with more than 5 observed clinician actions, as shown by the colors in the Clinician Probabilities plot.) (d) Probability of clinician actions for patients in the current state. (e) Description of the current patient state, as well as a chart showing most strongly contributing features according to SHAP. (f) Mortality rate of patients after being observed in this state.

senior clinicians in the ICU. APPs and fellows are generally less senior but still make independent decisions about sepsis treatments. Most participants were attending physicians, although their level of experience in the ICU varied significantly, as shown in Table 1. Sessions were conducted on Zoom and lasted between 20 and 50 minutes; participants received 50 USD in compensation.

During the study, participants used a simplified AI Clinician Explorer interface to assess and make treatment decisions for four patients while thinking aloud. Patients were selected from the MIMIC-IV dataset by the authors, as described in Sec. 4.1. Participants were free to explore all patient data prior to the time-step of interest, which included demographics, vital signs, lab values, mechanical ventilation settings, and a record of all treatments administered since the beginning of the patient's ICU stay.

Patients were presented in a randomized order. For each patient, participants saw a different version of the AI recommendation ("visualization condition"), presented in a fixed order with each successive condition containing more information. We elected to present visualization conditions in a fixed order to minimize cognitive burden on our participants and to give them an opportunity to progressively acquaint themselves with the features of the AI Clinician interface. As summarized in Fig. 3, the visualization conditions were as follows:

(1) **No AI.** Participants made the decision without an AI recommendation.
(2) **Text Only.** Participants were introduced to the AI and given a simple text-based recommendation (e.g., "For this patient, the AI recommends...")
(3) **Feature Explanation.** In addition to the textual recommendation, participants were shown a SHAP feature attribution chart explaining how the patient's state was determined (Fig. 3b).
(4) **Alternative Treatments.** Finally, for this condition participants were shown a bar chart with five possible treatment actions ranked by the AI-generated quality score. Bars were also color-coded by the frequency at which clinicians in the historical dataset took each action for similar patients, providing participants with a sense of both how common a decision was and the quantity of data that the recommendation was based on (Fig. 3c)[5].

The AI was introduced to participants as "Sepsis-AI," a tool that "analyzes patients' electronic health records and uses an artificial intelligence-based algorithm to recommend fluids and vasopressor doses that optimize mortality based on historical data." To prevent

[5]Although this visualization includes two different types of information (AI-predicted value and aggregate clinician behavior), we opted to include it as a single experimental condition to minimize the study burden for participants while collecting relevant think-aloud feedback for future iterative design.

Participant	Role	Years ICU Experience
P1	APP	1-2
P2	APP	3-5
P3	Fellow	3-5
P4	Fellow	1-2
P5	Fellow	<1
P6	APP	>10
P7	Attending	5-10
P8	Attending	5-10
P9	Attending	>10
P10	Attending	5-10
P11	Attending	>10
P12	Attending	>10
P13	APP	>10
P14	Attending	>10
P15	Attending	3-5
P16	Attending	>10
P17	Attending	>10
P18	Attending	5-10
P19	APP	>10
P20	APP	3-5
P21	Attending	3-5
P22	Attending	>10
P23	Attending	5-10
P24	Attending	3-5

Table 1: Summary of study participants, their roles, and their level of experience working in the ICU. ICU = intensive care unit; APP = Advanced Practice Provider.

bias we avoided referring to the AI as the "AI Clinician," since some participants may have been familiar with the discussion surrounding the original publication.

After reviewing each patient's history and current status, participants were asked to choose a treatment action to apply to the patient related to both the IV fluid amount and vasopressor dosage. Although the recommendation included specific dosages of IV fluids and vasopressors, we limited participants' choice set to (up to) three options in an effort to capture clinicians' first-order decision-making process, and to better replicate the way clinicians make resuscitation decisions at the bedside [56]. This design had the added benefit of increasing the analytic tractability of our results. The three options were: begin/increase, end/decrease, or leave unchanged. If a treatment strategy was not currently being used (e.g. patient not on vasopressor), the end/decrease option was removed, leaving participants with 4-6 possible actions per patient. After making each treatment decision, participants reported their confidence in their own treatment choice (on a 7-point Likert scale bounded by "not at all confident" and "extremely confident") and their beliefs about how challenging the case was (on a 7-point Likert scale bounded by "extremely easy" and "extremely challenging"). For all visualization conditions except for No AI, participants also rated the usefulness of the Sepsis-AI recommendation (on a 7-point Likert scale bounded by "not at all useful" and "extremely useful") and the degree to which the Sepsis-AI recommendation affected their confidence in their own treatment choice (on a 7-point Likert scale bounded by "... much less confident" and "... much more confident").

Once participants had entered their decisions on all four cases, we concluded the session with a brief semi-structured interview to understand how clinicians used the different visualizations as well as their perspectives on when they might perceive the AI to be helpful.

4.1 Case Selection

In any study involving acceptance of AI-generated recommendations, the scenarios that are chosen can have a large impact on participants' level of concordance with, and perceived trust in, the AI. As discussed in Sec. 3.1, a variety of factors make the "accuracy" of the AI Clinician's treatment recommendations impossible to determine with certainty. Therefore, instead of choosing cases based on a target level of accuracy, we deliberately chose cases and decision points in which the AI Clinician's recommendation was substantially *different* from historical clinician actions. Specifically, we used the AI Clinician Explorer to identify patients (and timesteps within patients) in which the AI Clinician recommended one treatment strategy (e.g. vasopressors and no IV fluids) for patients in a particular state, but a plurality of clinicians gave an alternative treatment (e.g. IV fluids and no vasopressors). This approach had the added benefit of replicating situations in which the AI recommendations might challenge clinician judgment. To make the cases more realistic, each patient was given a randomly-generated name, and the visualization was accompanied by a hypothetical clinical vignette summarizing the patient's status. The vignettes were written to provide only generic clinical context (e.g. "sepsis from a urinary tract infection"), with no information that could guide treatment decisions beyond what was included in the dataset. A summary of the cases is shown in Table 2.

4.2 Analysis

Ratings of confidence and AI usefulness were compared quantitatively to assess participants' attitudes towards each of the visualization conditions. Using the Python `statsmodels` package[6], ordinary least squares (OLS) regression models were fit to each 7-point Likert scale outcome using the visualization condition as the only predictor. Models controlling for the participants' role, gender, and years of experience yielded similar results. All models cluster standard errors at the respondent level using robust Huber-White estimators. For post hoc (pairwise) comparisons, we adjust for multiple tests using the Holm–Bonferroni method.

In the absence of a ground truth correct decision, treatment choices were evaluated in terms of their *concordance* against three reference standards for each patient: (1) the AI Clinician's recommendation, (2) the action taken by the clinician(s) on the actual patient in the MIMIC-IV database, and (3) the majority action chosen by attending physician participants in the No AI visualization condition. The latter served as an approximation for the "clinical consensus" decision for each patient, although (as expected) variability was observed even within these experts' decisions. To understand the relationship between visualization condition and

[6] https://www.statsmodels.org

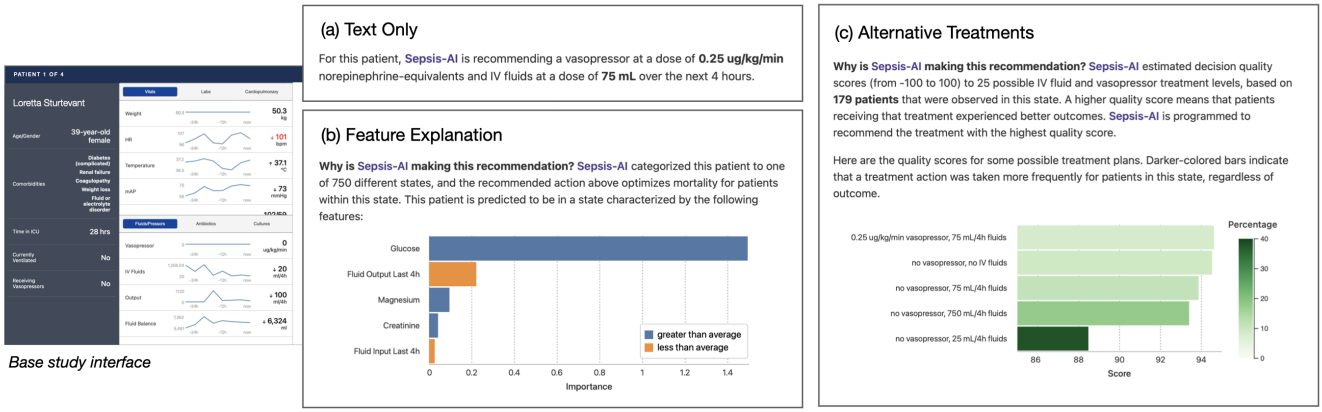

Figure 3: Visualization conditions used in the study. All participants were shown the base interface with patient trajectory views from the AI Clinician Explorer. The right half of the interface contained one of the following conditions when the AI was shown: (a) A textual description of the AI's recommendation, introduced in the Text Only condition and shown alongside subsequent visualization conditions as well. (b) The Feature Explanation chart shows the five variables that contributed most strongly to the AI's characterization of the patient state, and how each variable's values deviate from the average. (c) The Alternative Treatments chart shows five possible actions and the frequency at which clinicians historically took each action.

Patient Pseudonym	Ruth Silva	Loretta Sturtevant	Jeffrey Williams	Victoria Thompson
Demographics	76 y/o female	39 y/o female	74 y/o male	63 y/o female
Key Characteristics	mechanically ventilated, undiagnosed sepsis, currently hypotensive	type I diabetes, chronic renal insufficiency, previously received IV fluids	congestive heart failure, mechanically ventilated, recent admission, currently on high dose vasopressor	previously received vasopressor and IV fluids, currently hypotensive
AI Recommendation	no change in fluids increase pressors	increase fluids increase pressors	increase fluids decrease pressors	increase fluids no change in pressors
Original Clinician Decision	increase fluids no change in pressors	increase fluids no change in pressors	increase fluids decrease pressors	no change in fluids increase pressors
Majority Attending Decision	increase fluids no change in pressors	increase fluids no change in pressors	increase fluids decrease pressors	increase fluids no change in pressors

Table 2: Summary of the four patient cases selected for the think-aloud study. De-identified patient data was derived from the MIMIC-IV dataset. Three reference treatment decisions are shown for the time interval at which patients were presented: the AI Clinician's recommendation, the decision that was made by the clinician that treated the actual patient in the MIMIC-IV dataset, and the decision taken by the majority of attending physicians in our study in the No AI condition.

concordance, we used logistic regression in a similar fashion as above.

The 12.6 total hours of think-aloud sessions were machine transcribed using Descript[7] and manually cleaned in preparation for qualitative analysis. After reviewing these transcripts and notes in an interpretation session, the team developed a set of 23 codes that could systematically capture distinct decision-making behaviors. The four segments corresponding to each patient case were excerpted from each transcript and coded using this code book by two members of the research team. These coders met to discuss and resolve coding discrepancies and refine code definitions as needed. Finally, participants' broader viewpoints on decision-making using the AI were extracted using open coding, and themes were identified using affinity diagramming.

5 RESULTS

The following sections provide first an overview of participants' attitudes towards the AI in each of the visualization conditions (Sec. 5.1), followed by the decision-making behavior patterns observed in the think-aloud transcripts that help explain participants' use of the AI (Sec. 5.2). We then describe how as expert decision-makers, participants interrogated the underlying assumptions of the AI we presented them with, and reflected on how it could better assist them (Sec. 5.3).

[7] https://www.descript.com

5.1 Perceptions of Decision-Making with AI and Explanations

Across several measures, participants' perceptions of the AI varied as a function of visualization condition. Participants reported that the AI was more useful and that it increased their confidence to a greater degree when participants saw one of the two *explanation* conditions (Feature Explanation or Alternative Treatments), relative to when they saw the Text Only recommendation. Below, we report quantitative findings for the four Likert-scale responses we measured alongside relevant qualitative responses that help contextualize the data. The full pattern of results is reported in Fig. 4.

Usefulness of the AI. Visualization condition was associated with significant differences in participants' ratings of the AI's usefulness ($F(2, 69) = 4.251$, $p = 0.03$). Participants rated the AI as being more useful in the Feature Explanation condition than in the Text Only condition ($\Delta = 0.83$, 95% CI $[0.24, 1.43]$, $p = 0.018$) and directionally more than in the Alternative Treatments condition ($\Delta = 0.75$, 95% CI $[-0.03, 1.53]$, $p = 0.12$).

Effect of AI on confidence. Similarly, Visualization condition affected how participants rated the AI's impact on their confidence ($F(2, 69) = 7.946$, $p = 0.002$). Participants reported that the AI had a more positive effect on their confidence in the Feature Explanation condition than in the Text Only condition ($\Delta = 1.08$, 95% CI $[0.51, 1.66]$, $p < 0.001$) and directionally more than in the Alternative Treatments condition ($\Delta = 0.67$, 95% CI $[-0.05, 1.38]$, $p = 0.13$).

In the think-aloud sessions, several participants mentioned the positive effects of seeing explanatory evidence, either in the form of the Feature Explanation or Alternative Treatments. In the latter condition, clinicians particularly appreciated the AI's ability to compare outcomes of multiple possible decisions (P16, P18, P24): *"Seeing the different outcomes to those decisions in a similar case, I think is... the most convincing to change your clinical decision making"* (P24). However, the ability to see other clinicians' actions in this condition was less uniformly endorsed. Some respondents appreciated the additional reassurance of the sensibility of their decisions (P5, P7, P8, P24), while others (P10, P12, P17) expressed concern that it would steer novice clinicians towards common errors committed by less experienced clinicians: *"I'm highly suspect of what other people do. And I don't think that that's a good way to practice medicine"* (P17).

Confidence in treatment choice. Participants' confidence in their treatment choices was not significantly different across Visualization conditions ($F(3, 92) = 2.220$, $p = 0.11$) and no pairwise comparisons between conditions were statistically meaningful after adjusting for multiple comparisons (ps > 0.17). However, there was a directional increase in confidence ratings when explanatory visualizations were provided, particularly in the Alternative Treatments condition. We hypothesize that one benefit of the Alternative Treatments condition on decision confidence may have been that it presented evidence for multiple treatment options, not just the often-discordant top recommendation.

Perception of case difficulty. Visualization condition significantly affected perceptions of case difficulty ($F(3, 92) = 4.112$, $p = 0.02$), with the provision of AI and its associated explanations increasing perceived difficulty. Comparing individual conditions, we find that participants perceived the cases as being significantly less challenging in the No AI condition than in the Alternative Treatments condition ($\Delta = 1.08$, 95% CI $[0.46, 1.71]$, $p = 0.003$) and directionally less challenging than in the Feature Explanation condition ($\Delta = 0.79$, 95% CI $[0.14, 1.45]$, $p = 0.09$). Our interpretation of this pattern is that explanatory evidence may have prompted clinicians to consider more factors when making their decision, especially when explanations did not align with their mental model of the patient or when the recommendations went against their clinical judgment (P12, P17, P22, P24). The resulting cognitive burden may have made the case seem more difficult. For instance, one clinician noted:

> *"I would not have guessed that the decision or the recommendation was being based on something like a BUN [blood urea nitrogen] change. I assumed it was based on the CVP [central venous pressure], and I don't think that CVP was considered in [the Feature Explanation chart]. And so it kind of makes you try and guess where the recommendations are coming from, and you spend a little bit more mental energy thinking about that."* (P17)

5.2 Patterns of Interaction with the AI

In contrast to the attitudinal metrics, participants' actual decisions for each patient did not vary meaningfully as a function of visualization condition. The light blue bars in Fig. 5a show that clinicians chose the same treatment choice as the AI about 42% of the time regardless of the visualization condition—only a slight increase over the 33% base rate of concordance without seeing the recommendation at all. If any concordance (same choice according to *either* fluids or vasopressors) is included, participants again have roughly similar rates of agreement with the AI, except for a slightly lower rate in the Text Only AI condition (Fig. 5b).

When the AI was shown, we did observe a slight reduction in concordance with actions taken by the clinician treating the original patient as well as the majority attending decision (Fig. 5c-f). Specifically, the average full concordance with the majority attending decision was 50% across the three AI conditions, compared to 63% in the No AI condition. This may indicate that participants were swayed to do something other than the "typical" clinician action when using the AI. Yet the actions they ultimately took did not perfectly align with the AI either: out of the 36 AI-assisted decisions in which the participant did not fully agree with attendings, only 6 decisions showed full concordance with the AI. Though not statistically significant by logistic regression modeling, these somewhat counter-intuitive relationships led us to hypothesize that individual-level variations could be contributing to the roughly-constant overall rate of concordance with the AI.

Therefore, to gain more granular insight into when participants chose to accept AI recommendations, we turned to the qualitative analysis of participants' think-aloud transcripts. As described in Sec. 4.2, we developed codes to capture whether and how participants engaged with the AI along various aspects of its recommendations, as well as the reasons they provided for accepting or rejecting the recommendation. Grouping together participants with similar codes revealed four distinct behavior patterns, each of which was associated with different degrees of reliance on the AI. The four

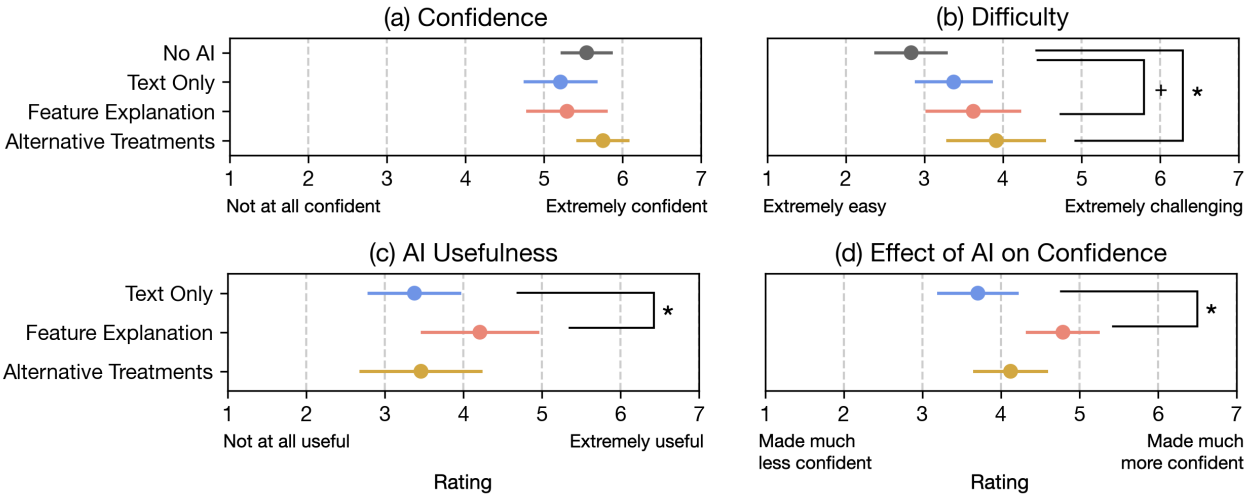

Figure 4: Summary of quantitative measures obtained from participants' self-ratings within each visualization condition: (a) participants' confidence in each decision; (b) how difficult they rated each case; (c) their rating of the usefulness of each version of the AI; and (d) how much the AI affected their confidence. * signifies $p < 0.05$; + signifies $p < 0.1$. Error bars indicate 95% confidence intervals.

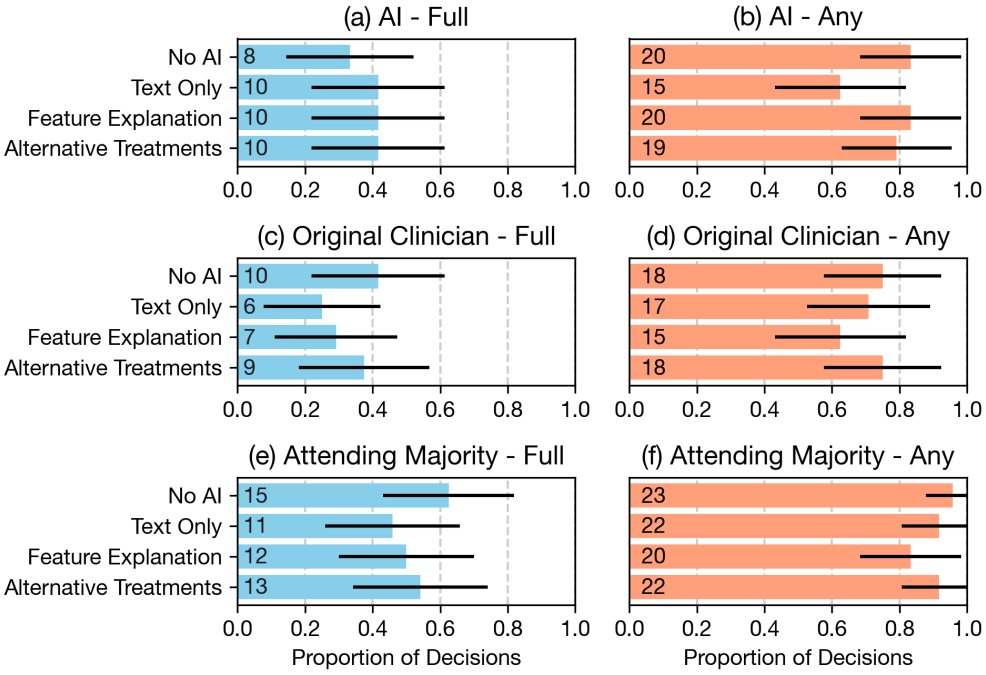

Figure 5: Rates of concordance between participants' decisions and three reference decisions: the AI recommendation, the decision of the clinician in the original dataset, and the decision taken by the majority of attending physicians in the No AI condition. The left column ("Full") shows agreement with both the IV fluid and vasopressor recommendations, while the right column ("Any") depicts agreement for *either* of the two treatment strategies. Each proportion is calculated over a total of 24 decisions; error bars indicate 95% confidence intervals.

behavior patterns are summarized in Fig. 6 and described in more detail below.

5.2.1 Ignore: Participant makes own decisions. For seven participants (21 total decisions using the AI), the AI never meaningfully influenced their decision in any way indicated by their think-aloud transcripts. Instead, their decision was predominantly driven by their initial clinical assessment, and not affected by recommendations even when explanatory visualizations were provided. These participants were often able to reject the recommendation because they were highly confident in their decision already, due to characteristics of the patient they identified as important based on their clinical experience: *"She's young and doesn't have heart problems and she's very net negative. So fluids would be the first thing I do for her, for sure"* (P5). Perhaps as a result of their confidence, these participants sometimes gave no verbal acknowledgement of the AI recommendation (3/21 decisions) despite the fact that it was clearly demarcated to them and they knew it was present. When participants did engage with the recommendation, they tended to critique it while holding their own assessment fixed. For example, P11 rejected a recommendation to give vasopressor and a small amount of IV fluid, arguing:

> *"She may be hypovolemic... because of the hyperglycemic state, but certainly I would not... start a pressor on this patient. [...] And IV fluids at a dose of 75 mLs over the next four hours... I disagree with that as well, because I think that this patient might be losing a lot of fluid on the urine output because of hyperglycemia.* (P11)

Interestingly, this engagement with the recommendation also affected some participants' confidence despite not affecting their decision. In these cases the AI served to either confirm the initial assessment—*"made me feel better about that decision" (P5)*—or, more commonly, to induce doubt when the recommendation was discordant (P6, P7, P17, P18). For instance, P17 noted that the recommendation *"to a certain degree made me question more than I would've. It actually probably made me think more about starting vasopressors, when any other time I would've just given the fluid bolus and not thought about it."* However, because these participants were already confident in their clinical reasoning and fairly settled on their decision, the AI recommendation was insufficient to cause them to change course.

5.2.2 Negotiate: Participant chooses aspects of the recommendation to accept. Unexpectedly, the most common behavior pattern we observed was of participants selectively adopting aspects of the recommendation as a form of auxiliary evidence. As with the Ignore group, the twelve participants in the Negotiate group still frequently made decisions that were not influenced by the AI (18/36 decisions). But in many cases, as shown in the middle columns of Fig. 6, they accepted at least one aspect of the recommendation:

(1) **Overall treatment choice.** In 14/36 decisions, participants agreed with the treatment recommendation for either fluids or vasopressors, but not both. For instance, P6 initially decided to follow a recommendation to begin vasopressors; however, upon re-examining the patient data, they noticed: *"She hasn't gotten any [fluids]... okay, interesting. Hmm. I would probably give a little fluid too."*
(2) **Quantity of treatment.** Participants engaged with the AI recommendation's specific dosage levels in 12/36 decisions, most often rejecting the values based on their knowledge of the patient: *"She's 39. I know she has chronic renal failure, but that doesn't mean that she cannot use fluid"* (P10). However, when the dosage values were within the range that participants would expect, they found value in the specificity of the AI recommendations:
    > *"I think a big challenge in the ICU is having a sense of... how much fluid to give a patient. [...] I think in that situation, I'm sort of more willing to give [the AI]... more of the nuance of the decision making. Like the big picture, we both seem to be in agreement. [...] And so then if the AI says, 'this is how much fluid I think they need in this period of time,' that's one less decision that I have to tax myself or burden myself with."* (P8)
(3) **Timing of treatments.** Finally, some participants expressed agreement with the AI's overall recommendation but refrained from making the recommended changes concurrently. For instance, P7 deferred the vasopressor component of one recommendation, reflecting that *"I don't necessarily disagree, it's a relatively small dose of norepinephrine. [...] I would probably start with the fluids, but then I would escalate to vasopressors if there was no response probably within a couple hours."* Conversely, one participant was swayed by a vasopressor recommendation to postpone their own decision to give fluids (P22).

We termed this behavior *negotiation* because participants assigned value or priority to various aspects of the recommendation, and thereby were able to arrive at an intermediate solution that balanced its most important aspects with their own intuition. Participants often prioritized parts of the recommendation using two factors:

(1) **Risk level and urgency.** In 12/36 decisions, participants used their perception of the severity of the patient's sepsis to decide how much to reconsider their treatment plan. Discordant recommendations for patients whose vitals seemed relatively stable were more likely to gain acceptance than those that appeared to be deteriorating. For example, weighing a recommendation to give vasopressors against their initial assessment to give fluids, P3 responded, *"I would say if I was alone without the computer helping me, I would give a trial of fluid. But I'm comfortable doing what they say. I think it's a coin toss anyway."*
(2) **Evidence presented by the AI.** More so than other groups, participants in this group used the explanatory visualizations as a source of evidence with which to understand the main point of the AI recommendation. For example, P18 was convinced by the Alternative Treatments chart to give fluids over their initial decision to start vasopressors:
    > *"Looks like very few people would have gone back on pressors, which is what I wanted to do. [...] I think it's fair. She needs something modestly aggressive because her [blood pressure] is quite low and it's been falling. [...] Yeah, I think this is a reasonable choice."*

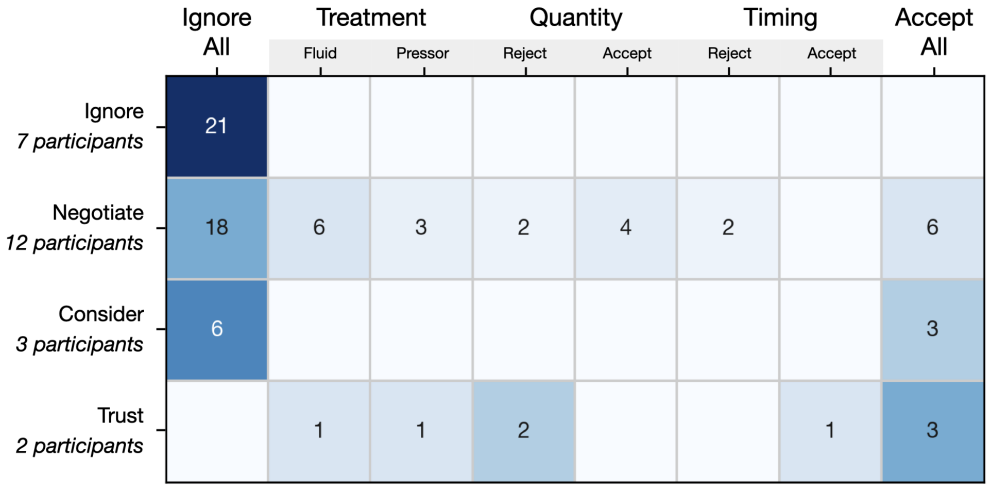

Figure 6: Patterns of reliance observed in participants' decisions, summarized from qualitative coding of think-aloud transcripts. The columns represent behaviors observed in an individual decision: Ignore All (decision not affected by AI), Treatment (accepted recommendation on one treatment strategy but not the other), Quantity (dosage levels), Timing (when to administer each treatment), and Accept All (fully changing the decision to align with the AI). The rows are sets of participants, grouped by these behaviors. Colors are normalized within participant groups (3 decisions per participant using the AI).

> Interestingly, this participant was discouraged from taking a less-common path by not only the AI's recommendation, but the summary of aggregate clinician behavior that the Alternative Treatments chart provided. On the other hand, negotiations sometimes led clinicians to ultimately reject the recommendation because they could not justify to themselves how an explanatory chart led to the recommendation. For instance, on a Feature Explanation chart, P24 questioned *"why a low [blood urea nitrogen] would lead to starting fluids and not vasopressors,"* ultimately leading them to go against the AI.

Perhaps because of the additional value they were able to obtain from the AI, Negotiate participants rated the recommendations more useful than other groups, with an average 7-point Likert rating of 4.6 (SD = 1.27) compared to 2.8 (SD = 1.79).

5.2.3 Consider: Participant conditionally accepts or ignores the recommendation. Three participants were similarly open to accepting the AI recommendation as the Negotiate group, but they either fully relied on the AI or made the decision on their own. Specifically, in 3/9 of their decisions, they yielded control of the decision to the AI, primarily based on their sense of uncertainty. For instance, P9 resolved to follow the AI recommendation for a difficult case: *"I am ambivalent about this one. Her [blood pressure] is slightly low. Her heart rate is actually coming down, fluid balance is positive... I think it's fine. We can do what the AI recommends."* Conversely, the same participant confidently dismissed a different recommendation: *"For this patient [the AI] is recommending a vasopressor dose of 0.25 of norepinephrine? Yeah, I don't think so."* In this way, the three Consider participants used the AI to drive their decisions when they were uncertain, but resumed control of decision-making in highly certain cases.

5.2.4 Trust: Participant always accepts some part of the recommendation. Finally, two participants were influenced by the AI for at least part of their decision in *every* decision they made. These participants often emphasized that the AI was based on objective data, perhaps leading them to consider its recommendations more willingly than other participants. For example, after reviewing the Alternative Treatments visualization, P16 reflected, *"This higher score means that they had better outcomes? Well then I'm gonna have to go with that. [...] The data looked pretty good."*

5.3 Perspectives on AI for Treatment Decision-Making

Throughout and after the think-aloud portion of each session, participants commented on how their decision-making processes in the simulated study environment compared with the decisions they made on a day-to-day basis. They also reflected on their habits and standard practices as clinicians, and how an AI might or might not be used to beneficially transform those practices. Below we discuss four themes that emerged from these discussions.

5.3.1 Participants' decisions are often guided by bedside information-gathering techniques rather than metrics used by the AI. We did not explicitly probe for next actions other than IV fluids and vasopressors, but eleven participants mentioned that a helpful next step would be additional data collection in the form of bedside assessment unavailable to the AI system. This usually took the form of dynamic assessments for fluid responsiveness via the physical exam, a procedure known as a "straight leg raise," or use of bedside ultrasound imaging. For instance, P21 noted that this information could help resolve a conflict with the AI on how much IV fluid to administer: *"If the AI was disagreeing with me, what I would do is walk into the room, do a leg raise, do a ultrasound... and then based*

on that information, I would decide how much volume to give." In fact, participants viewed this information as more reliable than any data used by the AI. They used this distinction to assert the superiority of human decision-making, reinforcing their identity as expert decision-makers while not outright rejecting the AI recommendation:

> *"At the bedside, I would acquire one piece or two pieces of reliable, better quality data than the algorithm has available. And then I would use that to make my decision [...] It's not fair to ask an algorithm to make a prediction that is as reliable as that is, because it doesn't have access to that."* (P23)

Participants similarly expressed concerns that the AI did not have access to more gestalt characteristics such as the patient's general appearance (P3, P7, P13): *"How ill do they look?"* To be clear, participants could not use these assessments during the study either, as they could only view the numerical data and general patient vignettes that we provided. Nevertheless, some clinicians (P20, P23) contrasted their confidence in these contextual assessments against the statistical nature of the AI: *"My bias as a clinician is that there is significant between-patient variability that is clinically significant, such that population level estimates used to inform individual patient care is fraught"* (P23).

5.3.2 The discretized dosage levels and time-scales used by the AI do not match with clinical practice. By design the AI Clinician collapsed all fluid and vasopressor dosage levels into 25 bins based on quantiles, ensuring a roughly uniform distribution of training labels. However, in practice this discretization led to confusion and doubt because all of the IV fluid bins were relatively low compared to the amounts clinicians were used to (presumably because most timesteps did not involve substantial fluid administration). For example, the third treatment level for fluids is 75 mL over four hours, to which one participant commented, *"I've never ordered such a small dose of fluids... To me that's like sprinkling water on her"* (P18).

The AI also aggregates data and provides recommendations at 4-hour intervals, which balances the rate of biometric data availability in the training dataset with the typical frequency of decision-making in the ICU. Clinicians overall found the 4-hour timescale appropriate for viewing the patient's trajectory and for making decisions on relatively stable patients, but they noted that they *"would not feel comfortable"* committing to higher-risk treatment decisions over that duration (P4, P7, P14, P17). Shorter-term decision points were viewed as a buffer against uncertainty about treatment responsiveness: *"In those situations where you're on the fence... you're gonna give your [IV fluid bolus], and you're gonna follow in that hour to two hours after they get the bolus to see if it had an effect"* (P14). In terms of measuring reliance on the AI, this reduction in timescale resolution led to clinicians effectively postponing agreement with the recommendation to a later decision (P4, P10, P17, P20), potentially nullifying the potential benefit of advance prediction by the AI.

5.3.3 Clinicians become skeptical of AI when it deviates from standardized or individual care practices. Participants often compared the AI's recommendations to the guideline-recommended practice of treating septic patients with hypotension, which comprises administering IV fluids (typically around 30 mLs per kilogram of body weight) and then vasopressors if the patient's blood pressure does not normalize [28]. These guidelines explicitly state that there is room for variation and that individual treatment plans should still be customized to each patient's unique circumstances, a fact acknowledged by participants—*"you have to sort of be willing to be flexible" (P16)*. Nevertheless, eight clinicians mentioned during decision-making that they would expect the AI to recapitulate rather than deviate from the guidelines. For instance, one participant voiced the tension they felt between the AI's recommendation and their training: *"So I see the score, but going off of the data and all of my knowledge of sepsis, we have to try to give her some fluids. We never jump straight to vasopressors"* (P19). It is impossible to know if the AI's recommendation to give vasopressors was a better decision, although some evidence shows that early administration of vasopressors could benefit patients [72] and expert opinion increasingly emphasizes vasopressors over fluid administration [44]. Yet these recommendations were dismissed as nonsensical given the patient's current status: *"Thinking she's not hypotensive. So why in the world is the AI asking me to start pressors? I'm rapidly losing faith in Sepsis AI"* (P12).

Participants also wanted the AI to concur with their personal practices, which they often defined in contrast to the predominant habits of other clinicians. For instance, two participants found the AI's recommendations *"a bit fluid aggressive"* (P2) at times, particularly because they perceived that many clinicians overuse fluids: *"I've seen it in ICU where we're just like bolusing them blindly. And the next thing you know, they're puffy like the Michelin man"* (P1). Five participants (particularly more experienced clinicians) framed their personal practices as the standard of comparison for both the AI and other clinicians, in that when *"the recommendation starts not very in line with what I would personally do with the patient, I don't think it's useful"* (P20). Because they viewed the AI as based on the actions of a general population of clinicians less skilled than themselves, participants were able to dismiss recommendations that aligned with norms they were already comfortable deviating from.

5.3.4 Rigorous and credible evidence of the AI's effect on outcomes is a prerequisite to trust. Aside from their reactions to individual decisions, several participants expressed that their overall level of trust in the AI would be determined based on the description of methodology and evidence provided to them before they ever used the tool (P3, P9, P12, P17). These participants believed they would read available background information on the tool, then either *"adopt it as a valuable tool or... shoot holes in it and say, 'I don't believe in this methodology and I'm not gonna use this tool anyway'"* (P17). The credibility of the AI would partially be determined by the reputability of its developers and the journal in which its validation study was published: *"If... there was a study in New England [Journal of Medicine] that said that Sepsis AI... improved outcomes, then I would say it could be kind of useful"* (P9). Once a high volume of credible evidence was available in favor of using the AI, participants believed they would more willingly trust its recommendations (P3, P9).

Although participants agreed that rigorous and credible evaluation was required, they were divided on how such a tool should

be evaluated. The most common suggestion was to conduct a randomized controlled trial with the AI to validate whether the second opinion it provided improved patient care (P12, P23); others suggested simply testing the association between recommendation acceptance and patient outcomes (P7, P9). In contrast, P17 suggested that the AI should simply use their decisions as the ground truth and aim to replicate them, as is currently done for diagnostic models:

> *"You could be convinced if somebody presented this to you and said, 'Hey, we've been looking at your clinical practice, and... you're 95% aligned with this. And so, you know, if we just set this to run, it's going to do the same thing that you would do 95% of the time, and you don't have to wake up.'" (P17)*

Regardless of what form the validation study took, participants agreed that upfront knowledge about the model's quality would not supersede clinical judgment on individual cases, leaving the door open to patterns of conditional and partial reliance even after trust is established.

6 DISCUSSION

We describe the development of an interactive CDS system for sepsis treatment, as well as a mixed-methods study that examined how clinicians interacted with that system to identify critical barriers to AI adoption in health care. Our results confirm prior findings suggesting that providing clinicians with explanatory evidence, either in the form of feature explanations or alternative treatment comparisons, can increase clinicians' perceptions of the AI's usefulness and confidence in their decisions [1, 80]. In terms of reliance on the AI, prior work studying reliance on CDS tools [12] and explainable AI [85, 90] led us to expect that clinicians would calibrate their own certainty against the AI and make a binary decision about whether to accept its advice in each case. However, only a few participants (the Consider group, Sec. 5.2.3) exhibited this dichotomous form of reliance. Instead, most participants engaged in a more nuanced form of partial reliance on the AI, often involving a negotiation between the initial clinical assessment and various aspects of the recommendation. Furthermore, several participants did not integrate the AI into their decision-making in any material way—for these participants the CDS only served to lower their confidence in their decision-making and increase the perceived difficulty of the case. Below, we discuss the implications of our results (key implications in bold) for the design of AI-based CDS and how to validate these systems in practice.

6.1 Designing AI for Complex Clinical Decisions

For a large number of health care decisions there is no evident "right answer" [23]. In these situations, successful AI should support clinicians in making better decisions on average, but must do so absent immediate feedback about the appropriateness of the recommendations. Out of the four broad decision-making behaviors we observed, the Negotiate behavior is closest to what one might consider an "appropriate" form of reliance on the AI in this setting. In contrast to the other three groups, participants who negotiated partial forms of reliance perceived a *range* of plausible next steps for each patient, not just a single action stemming from their clinical assessment. Furthermore, they were able to override aspects of the recommendation when they had specific contextual reasons to do so. On the other hand, clinicians had to develop their own assessments of which parts of the recommendation to rely on, perhaps resulting in more inconsistent decisions.

One approach to improve AI-assisted clinical decision-making could be to support negotiation by helping clinicians prioritize credible aspects of the recommendation. For instance, instead of recommending a rigid treatment plan over a four-hour interval, an algorithm could leverage historical data to compare the value of starting multiple treatments concurrently with the value of applying them sequentially, helping inform comparisons that clinicians may already be making. Alternatively, it could present evidence in favor of general treatment strategies at a binary level (e.g. fluids and no vasopressors) rather than specific values (e.g. 250 mL of fluids) unless the specific dosage was known to have an impact on mortality. These systems would serve to reinforce the belief that humans can make more nuanced decisions than AI systems, a belief we observed in this study. This type of AI would "know its limits" but still be able to guide decision-making by providing a framework by which clinicians could inform their decisions, rather than providing only prescriptive recommendations that are easily rejected. Though technically non-trivial to develop, such an AI may yield advice that can be more easily and consistently assessed by clinicians.

Our study also examined the effects of model explainability, an ongoing area of debate in AI-based CDS research [2, 3, 32], on participants' perceptions and behaviors using the AI. Our findings are consistent with prior XAI research [12, 85, 90] showing that explanations are a helpful complement to AI predictions, but that explanations alone will not significantly impact reliance. In particular, we observed that while the Feature Explanation chart helped participants decide how much weight to place on the recommendation overall, it did not support their ability to assign value to individual recommendation components. On the other hand, the Alternative Treatments approach may have better supported negotiation behaviors by allowing the AI to *"present its findings"* (P4) across a range of options. While similar to multi-class prediction charts used in prior work on diagnosis models [40, 80], the fact that the actions depicted in our visualization were quantitative (i.e., specific dosage levels) may have yielded the additional benefit of helping participants understand the overall trend predicted by the model, and thereby negotiate intermediate solutions. **Future explainable treatment recommendation systems could extend the Alternative Treatments approach to facilitate more nuanced comparisons of different choices, such as by projecting future patient states and outcomes conditioned on different choice sequences.**

Another important finding for explainability is its potential effect on cognitive effort. AI is meant to improve the efficiency of clinical decision making, saving clinicians time and reducing workload. Yet we found that explainable AI has complex effects on cognitive effort, especially when clinicians must decompose every recommendation into aspects with differing levels of credibility. On one hand, participants in the Ignore group tended to lose confidence and waste time comprehending a recommendation that ultimately would not affect

their choice. On the other, the visually dense explanations may have served as a cognitive forcing function to consider previously-neglected options [11], as Negotiate participants sometimes did. These results might suggest that the visibility and complexity of the AI recommendations be adjusted based on the users' confidence or the discordance between their decision and the AI. However, many participants also believed that once the AI was trusted, they would want to review it for confirmation of all their decisions, echoing Kulesza et al.'s findings that complete explanations tend to help despite requiring increased cognitive effort [51]. **Further research is needed to understand the tradeoffs between providing confirming recommendations to build trust, and saving clinician effort on discordant but non-useful recommendations.**

One simple solution to improve AI acceptance could be to focus adoption efforts on novice clinicians that may lack confidence in their ability to independently make clinical decisions. However, contrary to prior work showing negative effects of task expertise and AI familiarity on acceptance of AI recommendations [7, 25, 31, 40], the behavior patterns we observed did not appear correlated to seniority or experience level. Two of the seven Ignore participants were not attending physicians, while both of the Trust participants *were* attendings. While some prior work has examined how clinician demographics affect their needs for adopting AI [15], our interviews suggested an additional factor to consider: many clinicians are already regularly exposed to decision rules and behavioral interventions derived from historical data and expert committees, and they often hold diverging beliefs about how this clinical advice should influence decision-making. Even with similar experience levels, clinicians express varying degrees of awareness (and skepticism) of how recommendations are generated [48, 87], but the effects of differences in these attitudes have yet to be examined. **A better understanding of these perspective differences, and how they relate to experience level, may lead to designs that better serve people reluctant to factor AI advice into their decisions.**

6.2 Validating that AI-Based Decision Support Improves Outcomes

Unlike much prior work on AI-assisted decision-making in health care [14, 40, 78, 87], this study (1) used a real model trained to optimize treatment decisions, (2) provided clinical experts with real de-identified patient data, and (3) utilized a think-aloud protocol to capture further nuance beyond a multiple-choice survey. Participants responded to this realism in turn by revealing a more complex picture of clinical decision-making with an AI, one that in many ways does not fit the structure imposed by the AI. They expressed treatment goals in terms of information gathering (rather than always focusing on outcomes), adjusted dosage levels based on the patient's perceived needs, and temporally rearranged parts of the recommendation to more closely align with their standard practices. While this flexibility may well be desirable and even necessary in real-world decision-making, it creates an inherent tension with attempts to measure the quality of a system: **the more realistically an AI tool is integrated into clinical decision-making, the harder it becomes to assess whether the tool improves outcomes using standard validation techniques.**

Yang et al. [87] and Amann et al. [2] described a "chicken-and-egg" problem in which clinicians will not adopt AI recommendations unless they are backed by a credible validation study—yet in order for a validation study to succeed, clinicians need to adopt the AI's recommendations. This is particularly important in light of developing policies on AI in health care, such as the recent guidance by the U.S. Food and Drug Administration that treatment decision support systems such as the AI Clinician should be regulated as medical devices [81]. But unless clinicians are obligated to use the AI as part of a randomized controlled trial, the AI's effectiveness in prospective validation will be confounded with clinicians' low level of trust in the system, resulting in a poor (and possibly over-optimistic) estimate of its performance in deployment. Compounding this challenge, our results suggest that binary acceptance or rejection of recommendations in the sepsis treatment context is not an accurate indicator of the AI's effect on decision-making. After all, participants often gave credence to the AI, yet they rarely followed its recommendation completely. In an *in situ* validation study, how would partial or delayed acceptance be measured and assessed? **Developing acceptance metrics that account for partial reliance behaviors or changing reliance over time may help investigators perform validation studies that better capture potentially beneficial effects of the AI beyond binary acceptance.**

Another source of complication in validating AI-based recommendation systems is that reliability may vary significantly across different patient subgroups, requiring the user to develop a mental model of the AI's error boundaries [4]. However, even when clinicians in our study negotiated with the AI, they tended to approach its recommendations with a fixed level of trust or skepticism; their level of credence was rarely affected by the type of patient they were treating. **We suggest that instead of counting on end users to develop mental models of the AI's reliability, AI developers can collaborate with domain experts to extract, deploy and validate specific AI behaviors.** In other words, rather than considering the AI as an *agent* whose advice needs to be evaluated across a wide range of clinical decisions, we propose to use AI as a *source of evidence* whose recommendations can be separately assessed for specific subtypes of patients and disease states. This type of human-AI collaborative process could still yield more individualized recommendations than clinical trials (which are often too costly to run for all patient groups of interest), yet it would be more straightforward to evaluate than an AI that attempted to optimize for *all* patients. These selectively-validated recommendations can then be introduced to clinicians in stages, building the credibility of the AI while minimizing the chance of unforeseen AI errors.

6.3 Study Limitations

Although showing participants real AI recommendations for real patients yielded a more nuanced picture of decision-making, it also may have skewed our observations toward the particularities of the cases and recommendations we selected. Our depictions of the patient cases were limited to the structured data available in the MIMIC-IV dataset, meaning they had access to roughly the same amount of information as the AI. Additionally, we were unable to incorporate more domain-specific explanation techniques, such as

explainable RL (XRL), since they would have required substantial changes to the previously validated AI Clinician model. As a result, the SHAP explanations we showed focused on only one part of the model (the state clustering), thus limiting their potential usefulness to end users. Future work should investigate whether using more transparent model architectures and RL-specific explanation strategies improves clinical utility over the visualizations we tested.

Our study design and recruiting strategy was primarily focused on obtaining a rich set of think-aloud data for every decision we observed. While this resulted in ample data for qualitative analysis, it also meant we were unable to assess the statistical significance of some of our quantitative results, particularly levels of concordance with the AI. In the future we plan to build on these results by conducting a similar study with a larger pool of participants, enabling us to more accurately estimate the effects of providing AI explanations. Importantly, the present work indicated a need for more granular ways to collect structured data about decisions, which will inform the design of subsequent survey instruments.

Finally, this study was conducted with clinicians at a renowned academic hospital system in the United States. As such, they were likely more familiar than the modal clinician with the idea of applying clinical protocols or AI tools to improve decision-making. However, it is not clear whether this familiarity would tend to make them more or less accepting of tools such as the AI Clinician. Further research in institutions that have been slower to adopt clinical decision support tools is needed to evaluate the generalizability of these findings in other settings. Regardless, the fact that we observed such variation even in a relatively advanced hospital setting indicates that there is much work to be done in improving the acceptability of AI to clinicians.

7 CONCLUSION

To our knowledge, this paper is one of the first to rigorously assess clinicians' interactions with a real AI system that predicts the effects of treatment strategies under uncertainty. This form of AI aims to complement human decision-makers by revealing previously-unseen patterns in historical outcomes, in contrast to deep learning models that are simply designed to save clinician effort by recapitulating human decision-making. While many clinicians in our study were generally receptive to the idea of AI support, the ones who found the AI Clinician most useful in practice were those who saw it as a source of additional evidence—a piece of data that could inform their decision alongside their assessment. Reshaping these AI tools as a source of individually-validated recommendations may be one way to clarify their intended use and to facilitate evaluation of their impacts on decisions in the process. Together with advances in human-centered algorithm design and more nuanced decision metrics, we envision this work as a step towards AI-driven prediction tools that foster a refined notion of "appropriate reliance."

ACKNOWLEDGMENTS

We thank Ziyang Guo, Claire Chen, and Medha Palavalli for contributions to the modeling and visualization code; Billie Davis for assistance with transcription; and Dr. Emily Brant, Alex Cabrera, Nur Yildirim, Dominik Moritz, and John Zimmerman for helpful discussions around the manuscript. We also thank the numerous clinicians who participated in pilots and study sessions. This work was supported by a research grant from the United States National Institutes of Health (R35HL144804), by a National Science Foundation Graduate Research Fellowship (DGE2140739), and by the Carnegie Mellon University Center of Machine Learning and Health.

REFERENCES

[1] Lamia Alam and Shane Mueller. 2021. Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Medical Informatics and Decision Making* 21, 1 (2021), 1–15. https://doi.org/10.1186/s12911-021-01542-6

[2] Julia Amann, Dennis Vetter, Stig Nikolaj Blomberg, Helle Collatz Christensen, Megan Coffee, Sara Gerke, Thomas K. Gilbert, Thilo Hagendorff, Sune Holm, Michelle Livne, Andy Spezzatti, Inga Strümke, Roberto V. Zicari, and Vince Istvan Madai. 2022. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health* 1, 2 (2022), e0000016. https://doi.org/10.1371/journal.pdig.0000016

[3] Laura Arbelaez Ossa, Georg Starke, Giorgia Lorenzini, Julia E. Vogt, David M. Shaw, and Bernice Simone Elger. 2022. Re-focusing explainability in medicine. *Digital Health* 8 (2022). https://doi.org/10.1177/20552076221074488

[4] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S. Weld, Walter S. Lasecki, and Eric Horvitz. 2019. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. *33rd AAAI Conference on Artificial Intelligence, AAAI 2019* (2019), 2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429

[5] Gagan Bansal, Tongshuang Wu, Joyce Zhou, F. O.K. Raymond, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel S. Weld. 2020. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. *arXiv* (2020). arXiv:2006.14779

[6] Ian J Barbash, Billie Davis, and Jeremy M Kahn. 2019. National performance on the Medicare SEP-1 sepsis quality measure. *Critical care medicine* 47, 8 (2019), 1026.

[7] Sarah Bayer, Henner Gimpel, and Moritz Markgraf. 2021. The role of domain expertise in trusting and following explainable AI decision support systems. *Journal of Decision Systems* 00, 00 (2021), 1–29. https://doi.org/10.1080/12460125.2021.1958505

[8] Melissa Beauchemin, Meghan T. Murray, Lillian Sung, Dawn L. Hershman, Chunhua Weng, and Rebecca Schnall. 2019. Clinical decision support for therapeutic decision-making in cancer: A systematic review. *Int J Med Inform* (2019), 139–148. https://doi.org/10.1016/j.ijmedinf.2019.07.019

[9] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. *Conference on Human Factors in Computing Systems - Proceedings* (2020), 1–12. https://doi.org/10.1145/3313831.3376718

[10] Timothy G Buchman, Steven Q Simpson, Kimberly L Sciarretta, Kristen P Finne, Nicole Sowers, Michael Collier, Saurabh Chavan, Ibijoke Oke, Meghan E Pennini, Aathira Santhosh, et al. 2020. Sepsis among medicare beneficiaries: 1. The burdens of sepsis, 2012–2018. *Critical care medicine* 48, 3 (2020), 276.

[11] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. 5, April (2021). https://doi.org/10.1145/3449287 arXiv:2102.09692

[12] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. *Conference on Healthcare Informatics* (2015). http://openaccess.city.ac.uk/1189/

[13] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. *Conference on Human Factors in Computing Systems - Proceedings* (2019), 1–14. arXiv:1902.02960

[14] Carrie J. Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": Uncovering the onboarding needs of medical practitioners for human–AI collaborative decision-making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019). https://doi.org/10.1145/3359206

[15] Francisco Maria Calisto, Nuno Nunes, and Jacinto C. Nascimento. 2022. Modeling adoption of intelligent agents in medical imaging. *International Journal of Human Computer Studies* 168, September (2022), 102922. https://doi.org/10.1016/j.ijhcs.2022.102922

[16] Francisco Maria Calisto, Carlos Santiago, Nuno Nunes, and Jacinto C. Nascimento. 2021. Introduction of human-centric AI assistant to aid radiologists for multimodal breast image classification. *International Journal of Human Computer Studies* 150, August 2020 (2021), 102607. https://doi.org/10.1016/j.ijhcs.2021.102607

[17] Maurizio Cecconi, Laura Evans, Mitchell Levy, and Andrew Rhodes. 2018. Sepsis and septic shock. *The Lancet* 392, 10141 (July 2018), 75–87. https://doi.org/10.1016/S0140-6736(18)30696-2 Publisher: Elsevier.

[18] Centers for Disease Control and Prevention. 2021. What is sepsis?

[19] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[20] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. *International Conference on Intelligent User Interfaces, Proceedings IUI* (2021), 307–317. https://doi.org/10.1145/3397481.3450644

[21] Claudia Caroline Dobler, Allison S. Morrow, and Celia C. Kamath. 2019. Clinicians' cognitive biases: A potential barrier to implementation of evidence-based clinical practice. *BMJ Evidence-Based Medicine* 24, 4 (2019), 137–140. https://doi.org/10.1136/bmjebm-2018-111074

[22] Mark H Ebell, Randi Sokol, Aaron Lee, Christopher Simons, and Jessica Early. 2017. How good is the evidence to support primary care practice? *BMJ Evidence-Based Medicine* (2017).

[23] David M. Eddy. 1984. Variations in Physician Practice: The Role of Uncertainty. *Health Affairs* 3, 2 (1984), 74–89. https://doi.org/10.1377/hlthaff.3.2.74 arXiv:https://doi.org/10.1377/hlthaff.3.2.74 PMID: 6469198.

[24] David M Eddy. 1990. Clinical decision making: from theory to practice. Anatomy of a decision. *Jama* 263, 3 (1990), 441–443.

[25] Upol Ehsan, Samir Passi, Q. Vera Liao, Larry Chan, I-Hsiang Lee, Michael Muller, and Mark O. Riedl. 2021. The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations. (2021). arXiv:2107.13509 http://arxiv.org/abs/2107.13509

[26] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Martina Mara, Marc Streit, Sandra Wachter, Andreas Riener, and Mark O. Riedl. 2021. Operationalizing Human-Centered Perspectives in Explainable AI. *Conference on Human Factors in Computing Systems - Proceedings* (2021). https://doi.org/10.1145/3411763.3441342

[27] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. 2021. Deep learning-enabled medical computer vision. *npj Digital Medicine* 4, 1 (2021), 1–9. https://doi.org/10.1038/s41746-020-00376-2

[28] Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M. Coopersmith, Craig French, Flávia R. MacHado, Lauralyn McIntyre, Marlies Ostermann, Hallie C. Prescott, Christa Schorr, Steven Simpson, W. Joost Wiersinga, Fayez Alshamsi, Derek C. Angus, Yaseen Arabi, Luciano Azevedo, Richard Beale, Gregory Beilman, Emilie Belley-Cote, Lisa Burry, Maurizio Cecconi, John Centofanti, Angel Coz Yataco, Jan De Waele, R. Phillip Dellinger, Kent Doi, Bin Du, Elisa Estenssoro, Ricard Ferrer, Charles Gomersall, Carol Hodgson, Morten Hylander Møller, Theodore Iwashyna, Shevin Jacob, Ruth Kleinpell, Michael Klompas, Younsuck Koh, Anand Kumar, Arthur Kwizera, Suzana Lobo, Henry Masur, Steven McGloughlin, Sangeeta Mehta, Yatin Mehta, Mervyn Mer, Mark Nunnally, Simon Oczkowski, Tiffany Osborn, Elizabeth Papathanassoglou, Anders Perner, Michael Puskarich, Jason Roberts, William Schweickert, Maureen Seckel, Jonathan Sevransky, Charles L. Sprung, Tobias Welte, Janice Zimmerman, and Mitchell Levy. 2021. *Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock 2021*. Vol. 49. E1063–E1143 pages. https://doi.org/10.1097/CCM.0000000000005337

[29] Joseph Futoma, Anthony Lin, Mark Sendak, Armando Bedoya, Meredith Clement, Cara O'Brien, and Katherine Heller. 2018. Learning to Treat Sepsis with Multi-Output Gaussian Process Deep Recurrent Q-Networks. *ICLR 2018 Conference Blind Submission* 2017 (2018), 1–10. https://openreview.net/pdf?id=SyxCqGbRZ

[30] Joseph Futoma, Morgan Simons, Trishan Panch, Finale Doshi-Velez, and Leo Anthony Celi. 2020. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health* 2, 9 (2020), e489–e492. https://doi.org/10.1016/S2589-7500(20)30186-2

[31] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J. Berkowitz, Eva Lermer, Joseph F. Coughlin, John V. Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *npj Digital Medicine* 4, 1 (2021). https://doi.org/10.1038/s41746-021-00385-9

[32] Marzyeh Ghassemi, Luke Oakden-Rayner, and Andrew L. Beam. 2021. The false hope of current approaches to explainable artificial intelligence in health care. *The Lancet Digital Health* 3, 11 (2021), e745–e750. https://doi.org/10.1016/S2589-7500(21)00208-9

[33] Marzyeh Ghassemi, Mahima Pushkarna, James Wexler, Jesse Johnson, and Paul Varghese. 2018. ClinicalVis: Supporting Clinical Task-Focused Design Evaluation. (2018). arXiv:1810.05798 http://arxiv.org/abs/1810.05798

[34] Jennifer C. Ginestra, Heather M. Giannini, William D. Schweickert, Laurie Meadows, Michael J. Lynch, Kimberly Pavan, Corey J. Chivers, Michael Draugelis, Patrick J. Donnelly, Barry D. Fuchs, and Craig A. Umscheid. 2019. Clinician Perception of a Machine Learning-Based Early Warning System Designed to Predict Severe Sepsis and Septic Shock. *Crit Care Med* 47, 11 (2019), 1–18. https://doi.org/10.1097/CCM.0000000000003803

[35] Katharine E Henry, Rachel Korn, Anirudh Sridharan, Robert C Linton, Catherine Groh, Tony Wang, and Albert Wu. 2022. Human – machine teaming is key to AI adoption : clinicians ' experiences with a deployed machine learning system. (2022), 1–6. https://doi.org/10.1038/s41746-022-00597-7

[36] Tina B Hershey and Jeremy M Kahn. 2017. State sepsis mandates-a new era for regulation of hospital quality. *N Engl J Med* 376, 24 (2017), 2311–2313.

[37] Sungsoo Ray Hong, Jessica Hullman, and Enrico Bertini. 2020. Human Factors in Model Interpretability: Industry Practices, Challenges, and Needs. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (2020), 1–26. https://doi.org/10.1145/3392878 arXiv:2004.11440

[38] Michael H Hooper, Lisa Weavind, Arthur P Wheeler, Supriya Srinivasa Gowda, Matthew W Semler, Rachel M Hayes, Daniel W Albert, Norment B Deane, Hui Nian, Janos L Mathe, Andras Nadas, Janos Sztipanovits, Anne Miller, and Todd W Rice. 2015. Randomized Trial of Automated, Electronic Monitoring to Facilitate Early Detection of Sepsis in the Intensive Care Unit Michael. *Crit Care Med* 40, 7 (2015), 2096–2101. https://doi.org/10.1097/CCM.0b013e318250a887.Randomized

[39] Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. 2021. Explainable Artificial Intelligence Approaches: A Survey. (2021), 1–14. arXiv:2101.09429 http://arxiv.org/abs/2101.09429

[40] Maia Jacobs, Melanie F. Pradier, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of the antidepressant selection. *Translational Psychiatry* 11, 1 (2021). https://doi.org/10.1038/s41398-021-01224-x

[41] Russell Jeter, Christopher Josef, Supreeth Shashikumar, and Shamim Nemati. 2019. Does the "Artificial Intelligence Clinician" learn optimal treatment strategies for sepsis in intensive care? *arXiv* (nov 2019). https://doi.org/10.1038/s41591-018-0213-5 arXiv:1902.03271

[42] A Johnson, L Bulgarelli, T Pollard, S Horng, L A Celi, and R Mark. 2020. MIMIC-IV (version 1.0).

[43] Barbara E. Jones, Dave S. Collingridge, Caroline G. Vines, Herman Post, John Holmen, Todd L. Allen, Peter Haug, Charlene R. Weir, and Nathan C. Dean. 2019. CDS in a learning health care system: Identifying physicians' reasons for rejection of best-practice recommendations in pneumonia through computerized clinical decision support. *Applied Clinical Informatics* 10, 1 (2019), 1–9. https://doi.org/10.1055/s-0038-1676587

[44] Mathieu Jozwiak, Olfa Hamzaoui, Xavier Monnet, and Jean Louis Teboul. 2018. Fluid resuscitation during early sepsis: A need for individualization. *Minerva Anestesiologica* 84, 8 (2018), 987–992. https://doi.org/10.23736/S0375-9393.18.12422-9

[45] Ekaterina Jussupow, Kai Spohrer, Armin Heinzl, and Joshua Gawlitza. 2020. Augmenting medical diagnosis decisions? An investigation into physicians' decision making process with artificial intelligence. *Information Systems Research : ISR* tba, March (2020).

[46] Annika Kaltenhauser, Verena Rheinstädter, Andreas Butz, and Dieter P. Wallach. 2020. You Have to Piece the Puzzle Together": Implications for designing decision support in intensive care. *DIS 2020 - Proceedings of the 2020 ACM Designing Interactive Systems Conference* (2020), 1509–1522. https://doi.org/10.1145/3357236.3395436

[47] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *Conference on Human Factors in Computing Systems - Proceedings*. arXiv:2204.02310 http://arxiv.org/abs/2204.02310

[48] Saif Khairat, David Marc, William Crosby, and Ali Al Sanousi. 2018. Reasons for physicians not adopting clinical decision support systems: Critical analysis. *JMIR Medical Informatics* 20, 4 (2018). https://doi.org/10.2196/medinform.8912

[49] Michael Klompas and Chanu Rhee. 2020. Current sepsis mandates are overly prescriptive, and some aspects may be harmful. *Critical care medicine* 48, 6 (2020), 890–893.

[50] Matthieu Komorowski, Leo A. Celi, Omar Badawi, Anthony C. Gordon, and A. Aldo Faisal. 2018. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine* 24, 11 (2018), 1716–1720. https://doi.org/10.1038/s41591-018-0213-5 arXiv:1902.03271

[51] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing, VL/HCC* (2013), 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

[52] Vivian Lai, Chacha Chen, Q. Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. 1, 1 (2021). arXiv:2112.11471 http://arxiv.org/abs/2112.11471

[53] Min Hun Lee, Daniel P. Siewiorek, and Asim Smailagic. 2021. A human-ai collaborative approach for clinical decision making on rehabilitation assessment. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. https://doi.org/10.1145/3411764.3445472

[54] Scott M. Lundberg and Su In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 2017-Decem,

Section 2 (2017), 4766–4775. arXiv:1705.07874

[55] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2020. Explainable reinforcement learning through a causal lens. *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence* (2020), 2493–2500. https://doi.org/10.1609/aaai.v34i03.5631 arXiv:1905.10958

[56] Jason N Mansoori, Brendan J Clark, Edward P Havranek, and Ivor S Douglas. 2022. The Impact of Choice Architecture on Sepsis Fluid Resuscitation Decisions: An Exploratory Survey-Based Study. *MDM policy & practice* 7, 1 (2022), 23814683221099454.

[57] Aniek F. Markus, Jan A. Kors, and Peter R. Rijnbeek. 2021. The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. *Journal of Biomedical Informatics* 113, July 2020 (2021), 103655. https://doi.org/10.1016/j.jbi.2020.103655 arXiv:2007.15911

[58] Michael Moor, Bastian Rieck, Max Horn, Catherine R Jutzeler, and Karsten Borgwardt. 2021. Early prediction of sepsis in the ICU using machine learning: a systematic review. *Frontiers in medicine* 8 (2021), 607952.

[59] Trishan Panch, Heather Mattie, and Leo Anthony Celi. 2019. The "inconvenient truth" about AI in healthcare. *npj Digital Medicine* 2, 1 (2019), 4–6. https://doi.org/10.1038/s41746-019-0155-4

[60] Sonali Parbhoo, Shalmali Joshi, and Finale Doshi-Velez. 2022. Generalizing Off-Policy Evaluation From a Causal Perspective For Sequential Decision-Making. September (2022), 1–12. arXiv:2201.08262 http://arxiv.org/abs/2201.08262

[61] Xuefeng Peng, Yi Ding, David Wihl, Omer Gottesman, Matthieu Komorowski, Li Wei H. Lehman, Andrew Ross, Aldo Faisal, and Finale Doshi-Velez. 2018. Improving Sepsis Treatment Strategies by Combining Deep and Kernel-Based Reinforcement Learning. *AMIA ... Annual Symposium proceedings. AMIA Symposium* 2018 (2018), 887–896. arXiv:1901.04670

[62] Erika Puiutta and Eric M.S.P. Veith. 2020. Explainable Reinforcement Learning: A Survey. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 12279 LNCS (2020), 77–95. https://doi.org/10.1007/978-3-030-57321-8_5 arXiv:2005.06247

[63] Gordon D Rubenfeld. 2001. Understanding why we agree on the evidence but disagree on the medicine. *Respiratory care* 46, 12 (2001), 1442–1449.

[64] Sudarsan Sadasivuni, Monjoy Saha, Neal Bhatia, Imon Banerjee, and Arindam Sanyal. 2022. Fusion of fully integrated analog machine learning classifier with electronic medical records for real-time prediction of sepsis onset. *Scientific reports* 12, 1 (2022), 1–11.

[65] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. Ambiguity-aware AI Assistants for Medical Data Analysis. (2020), 1–14. https://doi.org/10.1145/3313831.3376506

[66] Nicolas Scharowski, Sebastian A. C. Perrig, Nick von Felten, and Florian Brühlmann. 2022. Trust and Reliance in XAI – Distinguishing Between Attitudinal and Behavioral Measures. *CHI 2022: Workshop on Trust and Reliance in AI-Human Teams* 1, 1 (2022), 1–6. arXiv:2203.12318 http://arxiv.org/abs/2203.12318

[67] Tjeerd A.J. Schoonderwoerd, Wiard Jorritsma, Mark A. Neerincx, and Karel van den Bosch. 2021. Human-centered XAI: Developing design patterns for explanations of clinical decision support systems. *International Journal of Human Computer Studies* 154 (2021), 102684. https://doi.org/10.1016/j.ijhcs.2021.102684

[68] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The human body is a black box": Supporting clinical decision-making with deep learning. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), 99–109. https://doi.org/10.1145/3351095.3372827 arXiv:1911.08089

[69] Mark P. Sendak, William Ratliff, Dina Sarro, Elizabeth Alderton, Joseph Futoma, Michael Gao, Marshall Nichols, Mike Revoir, Faraz Yashar, Corinne Miller, Kelly Kester, Sahil Sandhu, Kristin Corey, Nathan Brajer, Christelle Tan, Anthony Lin, Tres Brown, Susan Engelbosch, Kevin Anstrom, Madeleine Clare Elish, Katherine Heller, Rebecca Donohoe, Jason Theiling, Eric Poon, Suresh Balu, Armando Bedoya, and Cara O'Brien. 2020. Real-world integration of a sepsis deep learning technology into routine clinical care: Implementation study. *JMIR Medical Informatics* 8, 7 (2020), 1–16. https://doi.org/10.2196/15182

[70] Nirav R Shah and Thomas H Lee. 2019. What AI means for doctors and doctoring. *NEJM Catalyst* 5, 5 (2019).

[71] Manu Shankar-Hari, Gary S. Phillips, Mitchell L. Levy, Christopher W. Seymour, Vincent X. Liu, Clifford S. Deutschman, Derek C. Angus, Gordon D. Rubenfeld, and Mervyn Singer. 2016. Developing a New Definition and Assessing New Clinical Criteria for Septic Shock. *JAMA* 315, 8 (2016), 775–787. https://doi.org/10.1001/jama.2016.0289

[72] Rui Shi, Olfa Hamzaoui, Nello De Vita, Xavier Monnet, and Jean-Louis Teboul. 2020. Vasopressors in septic shock: which, when, and how much? *Annals of Translational Medicine* 8, 12 (2020), 794–794. https://doi.org/10.21037/atm.2020.04.24

[73] Dylan Slack, Sophie Hilgard, Sameer Singh, and Himabindu Lakkaraju. 2021. Reliable Post hoc Explanations: Modeling Uncertainty in Explainability. *Advances in Neural Information Processing Systems* 12, NeurIPS (2021), 9391–9404. arXiv:2008.05030

[74] Elizabeth K Stevenson, Amanda R Rubenstein, Gregory T Radin, Renda Soylemez Wiener, and Allan J Walkey. 2014. Two decades of mortality trends among patients with severe sepsis: a comparative meta-analysis. *Critical care medicine* 42, 3 (2014), 625.

[75] Lisa Stoneking, Kurt Denninghoff, Lawrence DeLuca, Samuel M. Keim, and Benson Munger. 2011. Sepsis bundles and compliance with clinical guidelines. *Journal of Intensive Care Medicine* 26, 3 (2011), 172–182. https://doi.org/10.1177/0885066610387988

[76] Harini Suresh, Steven R. Gomez, Kevin K. Nam, and Arvind Satyanarayan. 2021. Beyond expertise and roles: A framework to characterize the stakeholders of interpretable machine learning and their needs. *Conference on Human Factors in Computing Systems - Proceedings* (2021). https://doi.org/10.1145/3411764.3445088 arXiv:2101.09824

[77] Andrew K Teng and Adam B Wilcox. 2020. A review of predictive analytics solutions for sepsis patients. *Applied clinical informatics* 11, 03 (2020), 387–398.

[78] Sana Tonekaboni, Shalmali Joshi, Melissa D. McCradden, and Anna Goldenberg. 2019. What clinicians want: Contextualizing explainable machine learning for clinical end use. *arXiv* Ml (2019), 1–21. arXiv:1905.05134

[79] Eric J. Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* 25, 1 (2019), 44–56. https://doi.org/10.1038/s41591-018-0300-7

[80] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. 2020. Human–computer collaboration for skin cancer recognition. *Nature Medicine* 26, 8 (2020), 1229–1234. https://doi.org/10.1038/s41591-020-0942-0

[81] U.S. Food and Drug Administration. 2022. *Clinical Decision Support Software*. Technical Report. 1–26 pages. https://www.fda.gov/media/109618/download

[82] Baptiste Vasey, Myura Nagendran, Bruce Campbell, David A. Clifton, Gary S. Collins, Spiros Denaxas, Alastair K. Denniston, Livia Faes, Bart Geerts, Mudathir Ibrahim, Xiaoxuan Liu, Bilal A. Mateen, Piyush Mathur, Melissa D. McCradden, Lauren Morgan, Johan Ordish, Campbell Rogers, Suchi Saria, Daniel S. W. Ting, Peter Watkinson, Wim Weber, Peter Wheatstone, Peter McCulloch, and DECIDE-AI Expert Group. 2022. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. *Nature Medicine* 28, May (2022), 924–933. https://doi.org/10.1038/s41591-022-01772-9

[83] Dakuo Wang, Liuping Wang, and Zhan Zhang. 2021. Brilliant ai doctor in rural clinics: Challenges in AI-powered clinical decision support system deployment. *Conference on Human Factors in Computing Systems - Proceedings* (2021). https://doi.org/10.1145/3411764.3445432

[84] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing theory-driven user-centric explainable AI. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery. https://doi.org/10.1145/3290605.3300831

[85] Xinru Wang and Ming Yin. 2020. Are explanations helpful? A comparative study of the Effects of Explanations in AI-assisted Decision Making. *Intelligent User Interfaces, IUI '21, April 14–17, 2021, College Station, TX, USA* (2020), 318–328.

[86] Kevin C Wilson and Holger J Schünemann. 2011. An appraisal of the evidence underlying performance measures for community-acquired pneumonia. *American journal of respiratory and critical care medicine* 183, 11 (2011), 1454–1462.

[87] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. *Conference on Human Factors in Computing Systems - Proceedings* (2019). https://doi.org/10.1145/3290605.3300468 arXiv:1904.09612

[88] Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. 2023. Reinforcement Learning in Healthcare: A Survey. *Comput. Surveys* 55, 1 (2023), 1–36. https://doi.org/10.1145/3477600 arXiv:1908.08796

[89] Chao Yu, Guoqi Ren, and Jiming Liu. 2019. Deep inverse reinforcement learning for sepsis treatment. *2019 IEEE International Conference on Healthcare Informatics, ICHI 2019* (2019), 31–33. https://doi.org/10.1109/ICHI.2019.8904645

[90] Yunfeng Zhang, Q. Vera Liao, and Rachel K.E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020), 295–305. https://doi.org/10.1145/3351095.3372852 arXiv:2001.02114

[91] Alexandra Zytek, Dongyu Liu, Rhema Vaithianathan, and Kalyan Veeramachaneni. 2021. Sibyl: Understanding and Addressing the Usability Challenges of Machine Learning In High-Stakes Decision Making. Ml (2021). https://doi.org/10.1109/tvcg.2021.3114864 arXiv:2103.02071