Rethinking Human-AI Collaboration in Complex Medical Decision Making: A Case Study in Sepsis Diagnosis

Shao Zhang
zhang.shao1@northeastern.edu
Northeastern University
Boston, Massachusetts, United States

Jianing Yu
jiani.yu@northeastern.edu
Northeastern University
Boston, Massachusetts, United States

Xuhai Xu
xoxu@mit.edu
Massachusetts Institute of Technology
Cambridge, Massachusetts, United States

Changchang Yin
yin.731@osu.edu
The Ohio State University
Columbus, Ohio, United States

Yuxuan Lu
lu.yuxuan@northeastern.edu
Northeastern University
Boston, Massachusetts, United States

Bingsheng Yao
yaob@rpi.edu
Rensselaer Polytechnic Institute
Troy, New York, United States

Melanie Tory
m.tory@northeastern.edu
Northeastern University
Portland, Maine, United States

Lace M. Padilla
l.padilla@northeastern.edu
Northeastern University
Boston, Massachusetts, United States

Jeffrey Caterino
Jeffrey.Caterino@osumc.edu
The Ohio State University Wexner Medical Center
Columbus, Ohio, United States

Ping Zhang*
zhang.10631@osu.edu
The Ohio State University
Columbus, Ohio, United States

Dakuo Wang*
d.wang@northeastern.edu
Northeastern University
Boston, Massachusetts, United States

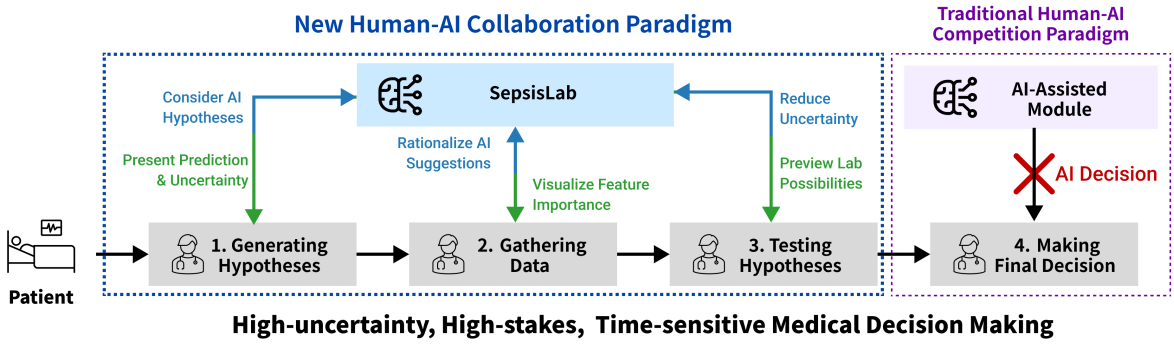

Figure 1: Existing EPIC Sepsis Module and Our Proposed Sepsis Decision-Support Module in Medical Decision Making Workflow. Our work focuses on sepsis diagnosis, a high-uncertainty, high-stakes, time-sensitive medical decision-making process. Physicians usually take four steps: (1) generating hypotheses, (2) gathering data, (3) testing hypotheses, and (4) making final decisions. Our study results point out that existing sepsis module is not helpful, forming a human-AI competition paradigm. We propose a new module to establish a human-AI collaboration paradigm.

ABSTRACT

Today's AI systems for medical decision support often succeed on benchmark datasets in research papers but fail in real-world deployment. This work focuses on the decision making of sepsis, an acute life-threatening systematic infection that requires an early diagnosis with high uncertainty from the clinician. Our aim is to explore the design requirements for AI systems that can support clinical experts in making better decisions for the early diagnosis of sepsis. The study begins with a formative study investigating why clinical experts abandon an existing AI-powered Sepsis predictive

*Correspondence to Ping Zhang and Dakuo Wang

This work is licensed under a Creative Commons Attribution International 4.0 License.

CHI '24, May 11–16, 2024, Honolulu, HI, USA

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642343

module in their electrical health record (EHR) system. We argue that a human-centered AI system needs to support human experts in the intermediate stages of a medical decision-making process (e.g., generating hypotheses or gathering data), instead of focusing only on the final decision. Therefore, we build SepsisLab based on a state-of-the-art AI algorithm and extend it to predict the future projection of sepsis development, visualize the prediction uncertainty, and propose actionable suggestions (i.e., which additional laboratory tests can be collected) to reduce such uncertainty. Through heuristic evaluation with six clinicians using our prototype system, we demonstrate that SepsisLab enables a promising human-AI collaboration paradigm for the future of AI-assisted sepsis diagnosis and other high-stakes medical decision making.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; • **Applied computing → Health informatics**.

KEYWORDS

Human-AI collaboration, Medical decision making, Sepsis diagnosis

ACM Reference Format:
Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M. Padilla, Jeffrey Caterino, Ping Zhang, and Dakuo Wang. 2024. Rethinking Human-AI Collaboration in Complex Medical Decision Making: A Case Study in Sepsis Diagnosis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3613904.3642343

1 INTRODUCTION

There is a growing interest from both academia and industry in the development of artificial intelligence (AI) to support medical decision making [10, 23, 31, 41, 46, 51, 91, 100]. Although the target scenarios may vary, from diagnostic decision making with medical imaging [83] to outpatient symptom triage [86], the ultimate goal remains the same: to reduce the burden of human medical experts while improving the quality of the final decision. Along this direction, many novel deep learning-based AI algorithms have been proposed, most of which yield promising predictive performance in their corresponding benchmark datasets [3, 14, 107], and some of them even outperform human experts in head-to-head competitions within controlled experimental settings [104].

However, the deployments of these AIs face more resistance in reality than their promising accuracy scores reported in research papers [64, 84, 88, 92, 93]. Luckily, more and more researchers have recently noticed the growing number of failure cases where AI-assisted medical decision-making systems are being abandoned by their target users. Recently, researchers have conducted various empirical studies to explore the cause of unsuccessful human-AI collaborative decision making [1, 12, 28, 47, 58, 65, 70, 78, 95, 96]. For example, human experts are the only ones responsible for an inaccurate diagnosis, while AI is not, so the clinicians trust their own judgment more than AI prediction [86, 90]. Based on these findings, they have proposed various suggestions for user interface and user experience design (for example, to improve physician adoption of AI with new eXplainable AI (XAI) features [1, 12, 28, 47, 58, 78, 105]).

In this work, we join the research effort to design AI to support medical decision making while focusing on the scenario of **sepsis diagnosis**. Sepsis is a common (48.9 million patients per year worldwide) yet life-threatening organ dysfunction triggered by a dysregulated response to infection [89]. The development of sepsis is very fast — without a timely diagnosis or proper treatment, a patient might die within a few hours from the initial onset of symptoms [103]. Compared to other medical decision-making scenarios (e.g., abnormal cell detection in medical imaging) where clinicians make a decision for that particular moment and with all the information they have at hand (e.g., cancer cells present or absent in the image), sepsis diagnosis is particularly challenging because: 1) clinicians need to decide not only whether the patient has sepsis at that moment, but also how likely this patient may develop sepsis in the near future (e.g., in a few hours); 2) they often do not have enough information to support their decision making. For example, the golden standard test for sepsis is the "blood culture" test in the Sepsis-3 guidelines [75]. However, it takes at least 8 hours to obtain the result, which would most likely be too late for a patient with sepsis [20, 26]. The early diagnosis of sepsis represents a common but under-explored decision-making scenario in the real world: it requires human experts with **specific domain knowledge** to cope with **high-uncertainty** and to make a **high-stakes** and **time-sensitive** decision based on **insufficient information**.

These special characteristics of sepsis diagnosis pose novel challenges for an AI system designed to support such decision making. There are some early efforts of research work that aim to design AI-based solutions to support sepsis diagnosis [11, 26, 29, 40, 49, 68, 75, 76, 97, 101, 104]. One notable effort is that a number of hospitals recently started to adopt a sepsis prediction and alerting module, **Epic Sepsis Module (ESM)**, in their existing Electrical Health Record (EHR) system [19]. The core of this sepsis module is a machine learning algorithm, which can take in a patient's EHR information and other biomarker data at that moment as input and predict a sepsis risk score as output [57]. When the predicted risk score is higher than a threshold, it suggests to human clinicians that a sepsis case presents, and human clinicians can agree, disagree, or dismiss such AI decision-making suggestions. Among the **four stages of the medical decision-making workflow** (*1. Generating hypotheses, 2. Gathering data, 3. Testing hypotheses, 4. Making decisions*) [79], this AI system is designed to provide support for the *final diagnosis* stage.

What do human experts think about such an AI-based sepsis diagnosis support system? Our work bridges this research gap with an interview study with six experienced clinicians who actively engage in sepsis diagnosis every day, and their hospital has recently adopted AI-based ESM technology. The results reveal that human experts believe the current AI module to be useless or even an intimidating "competitor" for the targeted high uncertainty, high-stakes, and time-sensitive decision-making scenario of sepsis early diagnosis (the bottom of Figure 1). Based on the findings, we designed **SepsisLab** with the goal of supporting the earlier stages of a medical decision making workflow for sepsis diagnosis (*hypotheses generation* and *data gathering for hypotheses testing*) instead of making a blunt prediction for the *final diagnosis decision*, as shown

in the top of Figure 1. Our SepsisLab system can predict and visualize the likelihood and uncertainty range of whether a patient has sepsis at the moment, and whether they may develop sepsis in the near future. In addition, SepsisLab can further suggest the most important but currently missing laboratory tests as an actionable suggestion for human clinicians, so that more data can be gathered to reduce uncertainty, leading to a more informed and higher quality final decision. A follow-up user evaluation study suggests that human experts appreciate our design, and they believe SepsisLab provides a better human-AI team experience compared to the existing human-AI competition paradigm. We envision our findings within the sepsis diagnosis example can shed light on the design of more human-AI collaboration paradigms for other domain-specific, uncertain, high-stakes, and time-sensitive decision-making tasks.

This paper demonstrates the following contributions:

- We conducted an empirical study to understand how human clinical experts interact with and perceive an existing AI-powered sepsis prediction module in their day-to-day work.
- We designed a new AI-assisted decision-making system, SepsisLab, which can predict and visualize the current and future likelihood and uncertainty of the onset of sepsis in a patient, and suggest actionable laboratory test recommendations to help human experts reduce the uncertainty of the final decision.
- We followed up with a user evaluation study, in which participants expressed a strong interest in adopting our system in their day-to-day work, and they believed our AI is no longer an "intimidating competitor" but more of a "collaborator".

2 BACKGROUND AND RELATED WORK

2.1 Sepsis Diagnosis

Sepsis is a severe, life-threatening condition that affects approximately 48.9 million patients worldwide, resulting in around 11 million sepsis-related deaths [11, 66, 89]. World Health Organization [89] highlighted the importance of early detection of sepsis symptoms and signs, along with the identification of biomarkers, for effective management. In modern clinical practice, the Sepsis-3 guidelines [75] serve as the gold standard for clinicians' diagnostic decisions. In the early diagnosis of sepsis, clinicians rely on clinical evaluations, laboratory test results, and blood cultures [26]. The diagnosis process represents a common, complex, but under-explored decision-making scenario: It is high-stakes (life-threatening), time-sensitive (a patient's severe symptoms developed only in a few hours), and highly uncertain (need extensive lab test outcomes that they often don't have at the moment of decision-making).

We contextualize the 4-step process in [79] for the current complex sepsis diagnosis workflow adopted by clinicians: (1) *Generating Hypotheses.* Physicians evaluate sepsis-risk patients and form hypotheses by the information from EHR system and physical examination but under significant uncertainty. (2) *Gathering Data.* Physicians order lab tests based on the most promising hypotheses to gather more information. (3) *Testing Hypotheses.* Based on lab test outcomes, physicians refine or expand their hypotheses. (4) *Making Decisions.* Physicians diagnose based on the revised hypotheses. As we show in Section 3.2, our formative study results validate these steps.

With the rapid growth in volume and diversity of Electronic Health Records (EHRs), AI-driven algorithms have been studied for the sepsis onset risk prediction task. Screening tools have been used clinically to recognize sepsis, such as quick Sequential (Sepsis-Related) Organ Failure Assessment (qSOFA) [75], Modified Early Warning Score (MEWS) [80], National Early Warning Score (NEWS) [77], and Systemic Inflammatory Response Syndrome (SIRS) [6]. However, those tools were designed to screen existing symptoms, as opposed to explicitly predicting sepsis prior to its onset, and their efficacy in sepsis diagnosis is limited. For example, prior studies show that qSOFA had low sensitivities in identifying sepsis in both prehospital and emergency department (ED) settings [21, 85]. In addition, deep-learning-based models are proposed to make sepsis onset predictions [48, 63, 67, 106]. Recent studies have employed attention mechanisms to explain models' inner workings [39, 104].

However, despite the advantages of AI models' performance, these methods still often fail to garner clinician confidence, thereby hindering their practical implementation in real-world clinical settings [24, 25, 45, 62]. For example, the Epic Sepsis Module (ESM) is the most widely used AI-based technique in current sepsis-related decision support practice [19]. The patient's data and lab test results (if any) are fed into the ESM module, generating a risk score. If the score is above a threshold, an alert will be sent to physicians and nurses, aiming to help with Step (4) in the current workflow. Yet a large number of studies have shown that its effect is impacted by a large number of external factors and cannot stably improve patient treatment effects [53, 88]. Existing HCI research on AI-CDSS in sepsis context primarily focuses on exploring treatment strategy choices during the treatment process [76] and investigating the transparency and explainability of AI algorithms [68], but do not delve into the challenges of decision-making during the diagnostic phase. Little is known about what human experts think about such an AI-based sepsis module in their diagnosis decision-making process. Our work aims to address this gap by interviewing experienced clinicians.

2.2 Challenges of AI-empowered Clinical Decision-Making Support

With AI's advancement, algorithms have been crafted to bolster clinical decision making, aiming to improve patient outcomes [76] and reduce clinician workload [34]. Numerous studies indicate that AI-supported clinical decision support systems (AI-CDSS) can effectively assist doctors. AI's suggestions can prompt doctors to reflect deeper on a patient's condition and alert them to potential disease progression [9, 43, 81]. For example, Yang et al. [95] noted that clinical order recommendation systems have garnered positive feedback, with doctors asserting that such recommendations enhance their work efficiency. Caballero-Ruiz et al. [8] quantitatively demonstrated that incorporating AI could diminish the time doctors spend evaluating patients.

However, a significant hurdle for AI-supported clinical decision making is liability. Given that clinicians bear the responsibility for medical decisions, they approach AI system predictions with utmost caution [5, 42, 86]. The opaque nature of AI algorithms makes it challenging for doctors to fully embrace AI's direct diagnostic and treatment suggestions [52, 65]. Recent research suggests

that embedding AI into EHR systems might inadvertently increase physician workload [5, 38]. Furthermore, a disconnection exists between what clinicians expect from AI and what AI actually delivers [10, 44, 81, 102]. Some research also found that current AI-assisted decision-making does not align with clinicians decision-making in depressive disorder [35] and type 2 diabetes [7] diagnosis process. Yang et al. [96] highlight that clinicians often adopt a 'wait and see' approach, seeking evidence to validate their hypotheses before deciding. In contrast, AI typically predicts outcomes based on available data, often failing to offer the evidence support that clinicians need.

In this paper, we delve deeper into clinicians' hands-on experiences and views on AI-CDSS within the current EHR system. Focusing on sepsis diagnosis, we aim to design a better form of collaboration between doctors and AI in current medical decision making to provide doctors with better decision support.

2.3 AI-supported Clinical Decision Support Systems Design

There has been growing research on designing AI-CDSS systems based on some general human-AI decision-making research [15]. Researchers design and implement a physician-facing interface and explore how physicians use the system to gain insights. Typically, AI-CDSS offers decision support to doctors by delivering predictions, risk evaluations, or suggestions [5, 9, 43, 68]. For instance, Yang et al. [95] introduce a system that auto-generates slides containing machine prognostics to aid clinicians in decision making. Some systems offer supplementary evidence or explanations to help doctors understand AI output, thereby enabling trust calibration. Examples include referencing literature [94], comparing with prior data Cai et al. [9], and bridging knowledge gaps [27]. Recently, some systems also support interactive interpretation aids, such as using attention mechanisms for model explanations [18] and enabling doctors to delve into nuanced concepts [9].

Previous research suggests that in sepsis diagnosis, the most important challenge is not to use XAI methods to explain the model outcome, but to use methods that are consistent with the doctor's cognition to establish trust between the doctor and the model [69, 76]. However, many current explainable AI methods cannot meet this goal for sepsis-related AI-CDSS [4, 22, 87]. In this paper, we further explore the challenges encountered by doctors in cooperation with AI-CDSS during the current early diagnosis of sepsis. We then design a better system to provide doctors with better diagnostic support.

3 FORMATIVE STUDY: CURRENT PRACTICES AND CHALLENGES OF AI-ASSISTED SEPSIS DIAGNOSIS

As discussed above, there is a research gap in how clinical experts perceive and interact with the AI-based module in their daily diagnosis workflow around sepsis patient cases. To better understand user needs and design challenges, we begin our project with an open-ended semi-structured interview [50] with six domain experts as a formative study to gather information on 1) the clinicians' daily practice of sepsis decision making, and 2) the user experience and user needs of AI-based sepsis decision support systems.

Table 1: Demographics of Physicians Participants. ICU - Intensive Care Unit, ER - Emergency Room, IM - Internal Medicine.

Participant ID	Gender	Departments	Job Title	Year of Experience
P1	Female	ICU	Staff Nurse	17 years
P2	Male	IM	Physician	14 years
P3	Male	IM	Physician	14 years
P4	Female	ER	Physician	4 years
P5	Male	ICU & ER	Physician	16 years
P6	Female	ER	Physician	10 years

3.1 Method

We recruited six clinicians who are domain experts and whose daily work involves decision making about sepsis diagnosis and also have used the Epic sepsis module (ESM). We used snowball sampling [30] to identify and recruit these participants by first reaching out to our colleagues and connections in related fields and then asking them to refer their connections. As shown in Table 1, all participants are active physicians or nurses working in departments (Intensive Care Unit, Emergency Department, or Internal Medicine) where clinicians most likely encounter sepsis patients. In the online prescreening survey, all participants reported that they "have used or are still using" the sepsis decision support module (ESM) in their EHR system (i.e., EPIC system). All interview sessions were conducted remotely via Zoom, and each interview session lasted, on average, 35 minutes. This study was pre-approved by the IRB committee of the first author's institution.

During the interviews, participants were asked to recall a recent case of sepsis encounter, detailing how the diagnosis decision of sepsis was made and what information and factors led to their final diagnostic decision, while not disclosing the patient's personal identifiable information (PII). Grounded in that aforementioned sepsis encounter experience, we prompted our participants to report further on their interaction and user experience with existing information technology (IT) systems, such as the EHR and ICU patient monitoring system. In particular, we asked them how they interact with and think about AI-driven ESM in their daily diagnostic process. We concluded our interview with their attitudes and user needs on the trend of deploying AI (not specific to EMS) to the medical decision-making process. The detailed semi-structured interview protocol can be found in Appendix A.

All interview sessions were audio-recorded and transcribed with the interviewees' consent. We employ an inductive approach [82], where two researchers in our team first independently coded the interview transcripts, then discussed and reconciled the coding schema, and finally reiterated and re-examined the data with the coding schema.

3.2 Result

We found that clinicians' decision-making process for sepsis diagnosis is high-stakes (life-threatening), time-sensitive (a patient's severe symptoms developed only in a few hours), and very uncertain (need extensive lab test outcomes that they often do not have at the moment of decision-making).

Clinicians walked us through the procedure of sepsis diagnosis, which provided a multifaceted journey from the patient's entry into the ER and (potentially) moving to the ICU or IM. They offered us more insights to enrich the 4-step procedure in Section 2.1. We summarized the workflow in Figure 2: (1) *Generating Hypotheses.* For a patient who has the risk of sepsis, physicians access and read a patient's vital signs and current state from the EHR system. They form a set of hypotheses given the patient's situation, yet these hypotheses are unclear with the large uncertainty. (2) *Gathering Information.* Based on the most promising hypotheses, they order specific lab tests to collect more information related to these hypotheses. (3) *Testing Hypotheses.* According to the lab test results, physicians narrow down or scale up their hypotheses. (4) *Making Action Decisions.* Based on the new hypotheses, physicians make decisions among three options: treating the patient, gathering more information, or withholding and waiting for new development of the disease.

All participants in general welcome the future of having more AI support for medical decision making, but, to our surprise, they strongly believe that the current Sepsis Module in EPIC (ESM) is not only useless, but also leads to additional meaningless work. Participants perceived the current ESM as a simple *sepsis risk prediction score*, and their dissatisfaction comes from (a) the risk prediction score is often **belated** (Section 3.2.1), (b) **inaccurate** (Section 3.2.2), (c) **no explanation** (Section 3.2.3), (d) **no actionable insights** (Section 3.2.4). In addition to these surface-level concerns regarding system design and algorithm performance, participants believe that the most fundamental issue is that the current paradigm of human-AI interaction design has a (e)**wrong focus** of AI assistance — it attempts to support the final decision of a complex medical decision-making process (i.e., high-stakes, time-sensitive and high-uncertainty). Together with the other issues mentioned above, human decision-makers feel challenged or even intimidated by the AI system, which eventually leads to all participants totally ignoring the current sepsis AI module (Section 3.2.5). We organize the results with these five aspects and dive into each of these issues.

3.2.1 Belated Sepsis Risk Prediction. All physicians complain that the current sepsis prediction is too late and thus useless in their decision-making process. This is due to the fact that sepsis is a life-threatening disease and can progress very fast. Thus, sepsis diagnosis requires human decision-makers to make a time-sensitive decision at the first encounter with the patients, despite they face huge uncertainty due to the lack of sufficient information about the patients. However, AI prediction must rely on data as input, but at first encounter, many of such data (e.g., vitals or lab results) do not exist or are not digitalized.

> *"So the big thing is that if we went strictly by the tool [AI-predicted sepsis cases] for our ED patients, we would be usually like three to four hours behind [human-diagnosed cases]."* (P6)

An extreme but quite common case reported by participants (P4, P6) is that shortly after human clinicians made the decision that a patient has sepsis, recorded the decision in EHR, and started to put in laboratory orders and antibiotic treatments in the EHR, the current sepsis risk prediction module consequently predicted that this is a sepsis case (see Figure 2 Step 4). Even worse, the current EHR system and hospital policy mandates the nurse or clinician to respond to this sepsis alert.

> *"If it triggers after two or three hours, I already know that, and I've already been treating them."* (P6)

Most participants explicitly demand an **early prediction** for early diagnosis of sepsis (P3, P4, P6), most likely at the first encounter in the emergency room (ER). They believe that such AI prediction could significantly speed up the following procedures and improve patient's final outcomes. However, the current sepsis prediction AI model cannot achieve that.

3.2.2 Inaccurate Sepsis Risk Prediction. The current algorithm design tends to have an extremely low sensitivity threshold at 13% (i.e., high false positive rate at 87%) to avoid missing any potential sepsis, because sepsis decision making has a patient's life at stake.

> *"the majority of times, I think it's inaccurate"* (P3)

As a result, participants (P3, P6) reported that they received an overly large volume of false alerts from the sepsis module due to the inaccurate sepsis risk prediction algorithm and the low sensitivity threshold.

> *"Any system that produces tons of alerts will induce alert fatigue, and people won't pay attention to it ... on average, there are already more than 20 interruptions per hour for an ER physician. So if you are adding more [inaccurate] interruptions to me, I'm not gonna pay attention, it becomes noise in the background."* (P6)

Even worse, due to the severe consequences of sepsis, the EHR system and hospital policy **forcefully mandated clinicians to manually verify** if they had taken appropriate diagnostic or treatment actions, or dismiss the alert but with a mandate note to explain to the system why the human clinician did not take any action (e.g., they have already taken sepsis treatment actions before the alert).

> *"The nurses get the BPA (sepsis risk alert) fired when it triggers that score. And then [nurses] have to put in [some notes] like an acknowledgment. A couple of options are like, "treatments already initiated", or "notified physician". Or you can silent it, I think, for 15 minutes or for half an hour, and then it'll fire again."*

3.2.3 Lack of Explanations. Participants (P1, P4, P5, P6) also mentioned that the current AI-predicted sepsis risk score is hard to interpret. Participants do not understand why an obvious sepsis case has a lower risk score than the score of a less obvious case.

> *"I don't think [another AI with higher prediction performance] would change anything because we were already unsure [how the current one works], and we were already looking for more explanations"* (P4)

They found that their multiple years of medical experience could not help interpret the relativeness of the score or the factors contributing to a score.

> *"On the Epic, there are too many parameters that I don't remember. [But] I know that it does not follow any of the other sepsis diagnosis criteria [being taught and used in practice] with the one, two, or three rating scale. And I know that it can go really high. And there are too many parameters."* (P3)

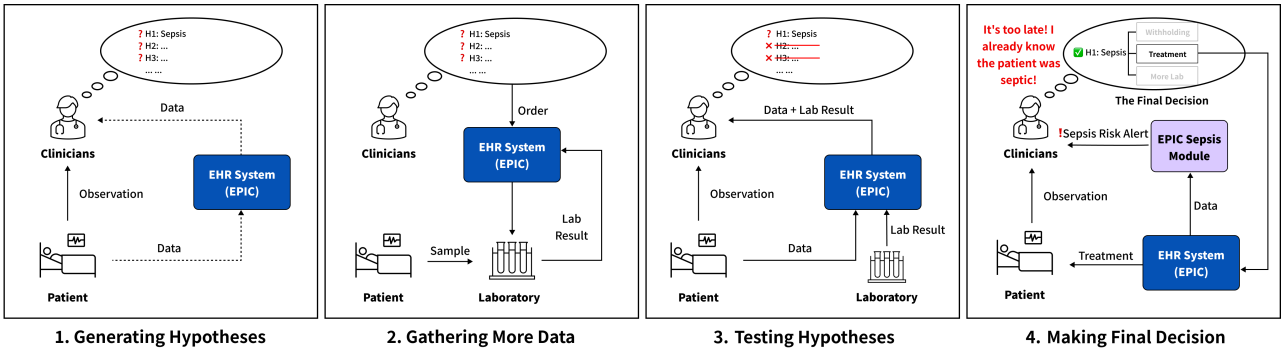

Figure 2: Existing Human-AI Interaction and "Competition" Paradigm. The current sepsis module mainly focuses on supporting the final decision-making stage [79], yet physicians often find the AI predictions are too late and not helpful.

Interestingly, the designer of the system already incorporated a feature importance score (a percentage of how much each factor contributes to the final prediction) as a simple explanation of the AI prediction, but they were hidden too deep in the interface. As a result, none of the participants except one (P4) was aware of its existence.

> *"I honestly never looked that deep into it (the risk score), I just see the color [of the risk score], and then I just go through my ["human"] algorithm that I've done for years and years."* (P2)

Even if such feature important percentage scores are at the top level, participants may still ignore them as there is already too much numerical information on the EPIC EHR interface.

> *"The good thing about EPIC and the worst thing about EPIC are the same — everything is there. It's kind of hard to figure out what you need to know. And it sometimes takes too many clicks."* (P3)

3.2.4 No Actionable Insights. Our interviewees also questioned the limited utility of it in their medical diagnostic process of sepsis. As we mentioned earlier, participants generally followed the four steps of medical decision making, from formulating hypotheses in their minds to finally making actionable diagnostic or treatment decisions. Most clinicians (P1, P2, P3, P4, P6) mentioned that they simply ignored the risk score during their diagnosis process because they were confused about the purpose of the risk score and did not know what to do as an actionable next step given a high or low risk score.

> *"So I'm not really sure what its goal is, but I can tell you that most of us ignore it (the sepsis risk score), because it has not proved helpful to what we do next".* (P4)

3.2.5 AI Helper or AI Challenger? The issues raised by participants in (Section 3.2.1, Section 3.2.2 and Section 3.2.3 may be addressed with a better algorithm or a more advanced user interface design. However, the concern of *lacking actionable insights* (Section 3.2.4) hinted at a more fundamental challenge that goes beyond the algorithm and interface design — the current AI-based sepsis prediction module mainly focuses on the prediction of the final decision outcomes, in our case, it is the sepsis risk score that suggests whether a patient has sepsis. Such output at the final decision stage implicitly "challenges" human decision-makers' authority and expertise in their roles of making that final decision.

> *"[The AI module] tries to just make a decision and tells me that [these patients] might have sepsis, so I have to do everything I'm supposed to do for [treating] them. That doesn't help."* (P5)

Due to the current prediction focusing too much on the final stage of decision making only, this human-AI interaction paradigm is essentially perceived as a human-AI competition, which challenges the human experts' expertise and intimidates their authorities and feelings.

> *"I think that AI can be very helpful as part of patient care, but I don't think it should replace the care and decisions a physician can make"* (P5)

Instead, participants believe that **AI can assist human experts in other places or stages of the medical diagnosis process**. For example, it can simply propose the sepsis possibility as a candidate hypothesis in the medical decision-making process.

> *"I would say that whether [AI] gave me like a 10% or a 90% sepsis risk score, I'm not sure that that would change my [decision]. If it [AI] simply tells me to think about it [sepsis possibility], then I'll just go to think about it."*

Alternatively, AI can suggest what kinds of laboratory data can be collected to support the test of the candidate hypothesis, from which physicians can obtain their desired actionable insights, and the uncertainty level can be reduced.

> *"We want to be better at knowing what to order and when and how to order it [lab or treatment]... in the current way that we're pushed by [the AI sepsis prediction score] right now is not useful, [and] it is correct most of times ... If a predictive model can trigger us to take an action [such as ordering lab or treatment] that prevents patients from getting sicker. That'd be amazing."* (P4)

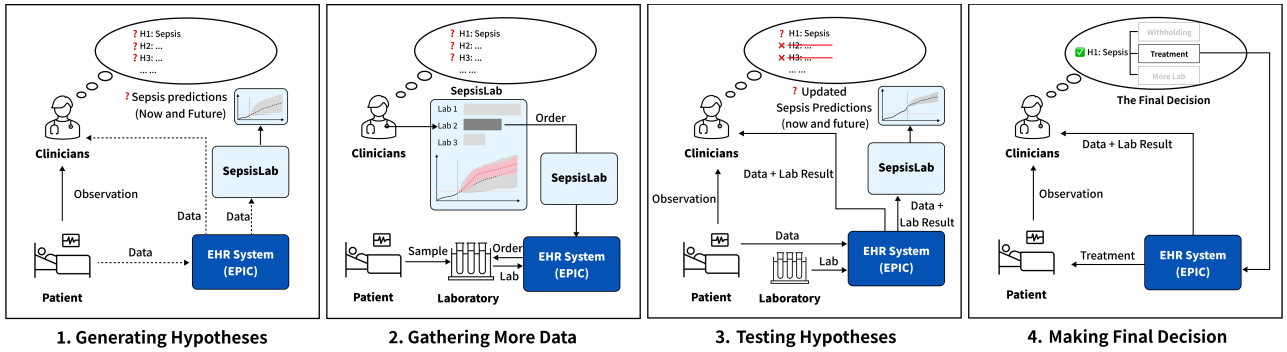

Figure 3: The Clinician's Medical Decision-Making Workflow with Support from SepsisLab. SepsisLab focuses on providing support to the intermediate steps of the clinical experts' decision-making process [79], as opposed to existing AI modules that focus only on the final decision-making stage. SepsisLab can generate predictions for the patient's sepsis onset possibility (as the risk score) now and in the future (Design Strategy 1, Design Strategy 4), as shown in Step 1; It can further suggest additional lab tests by their impact on model uncertainty (Design Strategy 2), and the interactive visualization can help clinicians select the most valuable lab tests to support their decision (Design Strategy 3, Design Strategy 4), as shown in Step 2; Once new data are collected, the prediction visualization will be updated (Step 3), helping clinicians test hypotheses. Then, following our Design Strategy 5, clinicians can generate new hypotheses or reach final decisions (Step 4).

3.3 Summary of Results

In summary, our formative study shows that the existing AI-driven sepsis risk prediction module does not support clinicians in their medical decision-making scenarios, because the current sepsis prediction algorithm is belated and inaccurate, the interface does not have explanations, and the AI prediction cannot be transformed into a diagnostic or treatment action. These challenges reveal a fundamental issue of the existing human-AI decision-making paradigm that human experts need AI to focus more on supporting their intermediate decision-making process, rather than predicting a final outcome. These findings shed light on our design of a new sepsis module with the goal of a new human-AI collaboration paradigm.

4 SEPSISLAB: A HUMAN-CENTERED AI SYSTEM TO SUPPORT EARLY DIAGNOSIS OF SEPSIS

In this section, we will start with the design strategies derived from the results of the formative study. Then, we will present both the user interface and the back-end algorithm of a novel human-centered AI system, SepsisLab, which aims to implement those design strategies to support clinical experts in making diagnostic decisions about sepsis. Our SepsisLab can predict the patient's current sepsis risk score, as well as the sepsis risk in the next 4 hours, based on patient history information and available vital signs and lab test values. Often times, some lab test results are missing but may also be critical for the diagnosis of sepsis. Therefore, our system can rank the top 5 lab tests that can reduce the uncertainty of the prediction and show them as recommendations to clinicians. Furthermore, our system has a counterfactual prediction module so that users can interactively review how each missing lab result may improve the prediction or reduce uncertainty before actually performing this lab test.

4.1 Design Strategies

Based on the stage 1 findings, we conclude five design strategies for the new design of the sepsis module.

Design Strategy 1: Performing Future Risk Score Prediction. As our formative study results suggest, clinicians do not need an inaccurate and even belated risk score prediction (Section 3.2.1 and 3.2.2). Instead, they need an accurate prediction score that is predicted ahead of time. This is also supported by previous empirical studies in sepsis diagnosis [32, 74, 88], which requires a better algorithm in the back-end of our system.

Design Strategy 2: Providing Accessible Model Explanation. Our interview results also suggest the need for an easily accessible section for sepsis risk prediction explanations (Section 3.2.3). This is a common issue found by previous research [69]. SepsisLab needs to have a simple design to present explanations in an easy-to-find and easy-to-understand manner.

Design Strategy 3: Revealing Actionable Insights and Suggestions. Moreover, a risk score, even predicted for future timestamps, cannot provide actionable insights, as suggested in Section 3.2.4. Clinicians base their diagnostic decisions on physical signs and lab test values from the patient, where gathering data (i.e., lab tests) plays an important role (recall Figure 1). Therefore, SepsisLab is designed to generate meaningful recommendations about potential lab tests.

Design Strategy 4: Displaying Uncertainty beyond the Risk Score. Sepsis diagnosis is a highly uncertain decision-making process. Providing a single risk score value may miss important information. Therefore, SepsisLab also calculates and displays the uncertainty in addition to the risk score. This is also aligned with the previous work about the advantages and benefits of XAI [87, 105].

Design Strategy 5: Shifting AI from Suggesting the Final Decision to Supporting Intermediate Stages. Finally and most importantly, the key takeaway from our formative study suggests

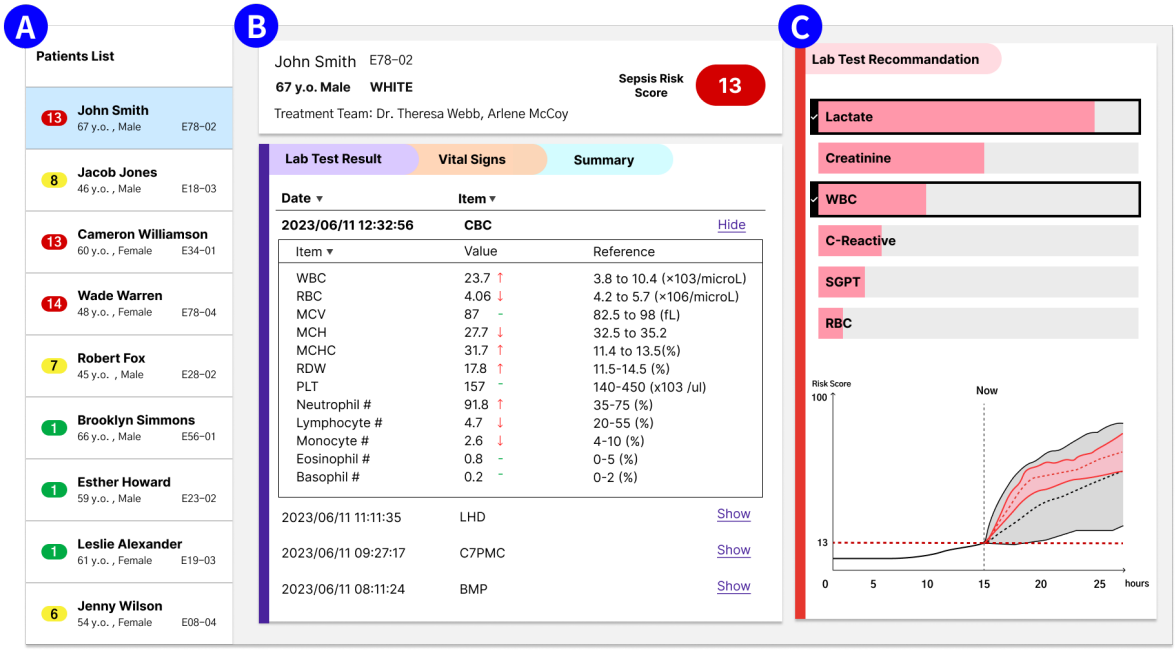

Figure 4: User Interface of Our Prototype System. (A) A list of patients with different sepsis risk prediction scores, colored from no risk as Green, to medium risk as Yellow, to high risk as Red. (B) The patient's demographics and the dashboard that includes the patient's vital signs, lab test results, and medical history. (C) Our SepsisLab system as an add-on to the existing EHR system. This UI currently illustrates that a clinical expert is examining a high-risk patient's data who was admitted 15 hours ago. The AI suggests the expert collect more lab results. The expert is interacting with the visualization to see if Lactate and WBC lab results were added, how the sepsis prediction and its uncertainty would change. All patient names and demographic information in this screen capture are random generated fake data for illustration purposes.

the need to shift AI's focus. Existing AI-based sepsis module mainly focuses on the final decision stage, creating a sense of competition for physicians and leading to the abortion of the module (Section 3.2.5). To address this challenge, SepsisLab is designed to support human experts' intermediate decision making stages, including generating hypotheses, gathering data, and testing hypotheses (Step 1 to 3 in Figure 1). In such a way, our system can build a new human-AI collaboration paradigm, where AI can actually team with experts to support what they need.

Combining these five design strategies, SepsisLab supports a new medical decision-making workflow for sepsis diagnosis, as shown in Figure 3. SepsisLab can generate predictions for a patient's sepsis risk score now and in the future, which can support the generating hypotheses stage. Moreover, it can further suggest additional lab tests that clinicians may gather to support their decisions. With the interactive visualization, our system helps them select the most valuable lab tests. This can provide actionable insights and support clinicians' data-gathering process. Once new lab test data are collected, the prediction visualization will be updated and assist clinicians in testing their hypotheses.

4.2 Front-End User Interface Design

Following the design strategies, we designed and implemented the new sepsis module based on human-AI interaction guidelines [2, 73]. The user interface (Figure 4 [1]) includes three parts: (A) Left: The current patient list, (B) Middle: The selected patient's demographic information, medical history, lab test results, and vital signs monitoring, (C) Right: Our AI-powered Lab Test Recommendation Module, SepsisLab, including Lab test recommendation, risk score predictions, and counterfactual explanation. We used de-identified patient data from MIMIC-III [37] as the data of the prototype system (data in Figure 4 B & C).

It is noteworthy that we deliberately designed the sepsis score of our predictive model's outcome to be the same as that of the current EPIC Sepsis Module. So that we could help SepsisLab have an easy integration into the existing EPIC system, where the clinicians can find it familiar and easy to use. We introduce each feature in Figure 4 C one by one.

Future Risk Prediction with Uncertainty Visualization. As mentioned in Section 4.1, one core part of SepsisLab is a predictive algorithm that generates future prediction of sepsis risk scores

[1]Note that we used MIMIC-III data for our algorithm illustration. All patient names and demographic information in this screenshot are random generated fake data for illustration purposes.

(**Design Strategy 1**). As shown in the bottom part of SepsisLab interface, we design a time-series plot to visualize both the historical (the solid line) and expected future (the dashed line) risk prediction trajectory over time.

Moreover, for each risk score prediction, the model also generates the uncertainty range of the expected value, as shown in the gray area in the line plot (**Design Strategy 4**). We selected visualized confidence intervals to display the prediction uncertainty, because prior work has found that confidence intervals evoke high levels of trust [59, 73]. While prior research has found that other uncertainty visualization techniques produce more accurate risk judgments [60], our visualization aims to show relative risk rather than for clinicians to read off specific values. Future work would be apt to consider the effects of various visualization design choices.

Feature Importance Visualization. The algorithm takes lab test item values as the input and generates risk score prediction. Each newly collected lab test item can be used to update the model and (potentially) reduce the model uncertainty. Therefore, we designed a ranked horizontal barplot on the top of the SepsisLab interface to visualize the important items that contribute to the prediction uncertainty reduction (**Design Strategy 2**). The item with the highest importance is ranked on top of the barplot.

Lab Tests Recommendations. Combining the two parts of future risk prediction and feature important visualization, we added the lab tests recommendation function into our interface (**Design Strategy 3**). As mentioned above, different newly collected lab items can change the model prediction uncertainty. Therefore, SepsisLab recommends an item list ranked by their importance. The clinician can select one or multiple lab items and observe how their test results could influence the model's risk prediction trajectory (the red dashed line) and the corresponding uncertainty range (the red area). Note that the red line and area are counterfactual values that are estimated by the algorithm (more details in Section 4.3).

SepsisLab supports the clinicians to interact with the interface. Figure 5 visualizes the interactive process by picking different potential lab test items. By comparing different combinations of the lab test items, the clinician can obtain a better understanding of the model and make the decision to order appropriate lab tests to collect the actual item values, which then truly update the model's prediction trajectory and uncertainty range. Overall, the interface follows Section 4.1 to support clinicians' intermediate decision making stages (**Design Strategy 5**).

4.3 Back-End Algorithms Design

To support the design strategies (Section 4.1) and the UI features (Section 4.2) informed by the formative study results, our back-end consists of three sub-modules: 1) an LSTM-based predictive model that can take only partial or little data of a patient as input and generate prediction scores of sepsis risk for the patient in the upcoming period as output; 2) a lab test recommendation module based on the uncertainty estimation from the previous predictive model; and, 3) a counterfactual generation module that can show the users how a hypothetical lab result may change the sepsis risk prediction scores and uncertainty range of the predictive model, and enables users to interact with the visualization chart in the front-end.

4.3.1 LSTM-based Predictive Model for Sepsis Risk Prediction. EHR data are typically a temporal sequence of patient activities in a hospital system. Depending on how frequently a patient visits a doctor or takes a lab test, the data sequence may be very sparse and irregular for a particular patient. Prior works [16, 17, 55, 99, 104] suggests Long Short-Term Memory (LSTM) [33], a special type of RNN model, has consistently demonstrated their remarkable performance in clinical risk prediction tasks using EHR data (see Table 3 from [104] in Appendix B as the evidence of the model performance). Recurrent Neural Network (RNN) model architectures are more suitable for the sepsis predictive tasks on the temporal observational data with irregular time intervals. In this study, we select Long Short-Term Memory (LSTM) [33] as the backbone of the prediction framework, which is able to capture both long-term and short-term clinical information in patients' EHR history, and thus improve clinical prediction performance.

To satisfy design strategy (1) that clinical experts want a prediction model that has a high accuracy and can predict ahead of time, we adopt the LSTM-based sepsis prediction model from [104] as our base model to support the prediction of the sepsis onset risk in the next 4 hours. We provide our LSTM model implementation details and parameters in Appendix C. The extracted static information vector (e.g., patient's demographic and history) is used to initialize the hidden state of LSTM. Then, the LSTM takes a sequence of collection data (e.g., vital signs and laboratory values) in addition to their occurring time as inputs and generates a sequence of the latent health state. Sometimes an observation may be missing (e.g., a patient has not performed a lab test or their previous lab test result has been outdated), thus a value embedding [98] is used to map the observations into vectors.

A variable attention module that can handle varying numbers of inputs is followed to generate a fixed-size vector that is sent to LSTM. With such a model design, the predictive model can start making predictions at the first encounter with the patient even there is not much data, which satisfies the **Design Strategy 1** that users want to see predictions into the future. The attention module can automatically focus on important variables, and the learned attention weights can be used to interpret the prediction results — this enables the users to see each input feature's importance score contributing to the predictive model — satisfying **Design Strategy 2**. After all the output vectors of LSTM are produced, a collection attention module is followed to combine the sequence of output vectors into a vector. Finally, a fully connected layer and a Sigmoid layer are followed to predict the sepsis onset probability.

4.3.2 Lab Test Recommendation Based on Uncertainty Estimation. The AI model's prediction always comes with certain degrees of uncertainty. In the sepsis early diagnosis scenario, a new patient when they just arrived ER may not have any lab test results in the EHR, thus many missing values as input to the predictive model. Due to this uncertainty, simply looking at the predicted sepsis risk score without the certainty level, users may not be able to accurately evaluate the trustworthiness of an AI prediction, and that is why participants reported that a patient with a high risk score for sepsis is not necessarily more accurate or more urgent than a patient with a low risk score.

Figure 5: The Interactive Lab Test Recommendation Module in SepsisLab. (a) The clinician can get an actionable lab item test recommendation list from SepsisLab. The items are ranked by their importance to reduce the uncertainty of the sepsis future prediction. (b) The clinician can interact with SepsisLab to select a lab item and see its expected influence of the lab test result on the model uncertainty via a counterfactual prediction. (c) The clinician can select multiple lab items and see their combined expected influence of the results on the uncertainty.

We estimate uncertainty and reduce uncertainty via improving the aforementioned LSTM-based predictive model: We hypothesize that the missing variables follow a Gaussian distribution so that we can estimate the parameters (i.e., mean and covariance) for each missing variable. Based on previous work that has shown superior performance in missing value imputation with deep learning [98], we adopt Monte Carlo Simulation (MCS) to sample the missing values many times and compute the uncertainty with the standard deviation of the outputs with MCS.

Two thresholds th_s and th_e ($1 > th_s, th_e > 0$) are set according to the desired sensitivity and precision to decide whether new laboratory values should be requested. If the output sepsis onset probability $p > th_s$, the model predicts that the patient will have sepsis onset after 4 hours. Sometimes, the model may be uncertain about the prediction results. We define the **uncertainty** as Shannon entropy [71]: $e = -p\log(p) - (1-p)\log(1-p)$. If the uncertainty $e > th_e$, the model will **recommend clinicians to collect more clinical variables**, for example laboratory values. Then, the model takes the updated values as input and can output new results with higher confidence.

To confirm that the recommendation can improve the model performance in the absence of certain laboratory values, we conducted experimental tests on the MIMIC-III dataset [37]. As shown in Table 2, in our implementation, with an LSTM model, our recommendation algorithm performs comparable to full observation setting models and outperforms the masked setting models by approximately 10% with only 9.6% extra laboratory values requested by our recommendation algorithm. The results indicate that our recommendation algorithm can achieve performance nearly equivalent to that under full observation, thus enabling accurate predictions even with fewer lab test results, without compromising prediction precision. We provide our recommendation algorithm implementation details including model parameters in Appendix C.

This algorithm design satisfies both **Design Strategy 3** and **Design Strategy 4**. The model focuses on recommending missing laboratory values to the clinician so that the clinician perceives such recommendations as an actionable suggestion. Additionally, the uncertainty estimation and visualization shift users' attention from the accuracy of the AI-predicted final decision, but to the reduction of such uncertainty in decision making.

4.3.3 Counterfactual Prediction to Explore Uncertainty Reduction Without the Cost of Performing a Lab Test. From AI's perspective, it would love clinicians to perform all kinds of lab tests on a patient so that it can reduce most of its uncertainty and predict at a higher accuracy. However, from the human's perspective, it is too costly and inhumane. While the lab test recommendation algorithm can identify the most informative laboratory test that is missing and reduce uncertainty of the prediction, it's still hard for doctors to intuitively see the value of these labtests. Therefore, to further

Table 2: Improvement of Our Recommendation Algorithm on AUC on MIMIC-III. Masked: all lab test results are deleted (simulating a patient just arriving at the ER). Our Algorithm: 9.6% lab test results are actively selected and repeated (simulating the clinician carefully selected most critical lab tests to order for the patient and gradually adding more labs if necessary). Full-observed: all the lab test results are used for prediction (assuming a patient has been hospitalized for a long time, so they have done many labs). The results show that our recommendation algorithm performs comparable to full observation models and outperforms the masked setting models by about 10% with only 9.6% extra laboratory values requested. The improved column denotes the improved performance with our recommendation algorithm, compared to the results in the masked setting.

Method	Masked	Our Recommendation Algorithm	Full-observed
Logistic regression	0.72	0.78	0.79
Random forest	0.75	0.81	0.83
Gradient boosting trees	0.75	0.83	0.85
LSTM [104]	**0.79**	**0.89**	**0.90**

help doctors make decisions we design a mechanism that the algorithm can also output counterfactual predictions to show **how much uncertainty can be reduced** without actually performing and collecting the recommended laboratory test results, which can provide more information in the process. For each variable i, we first sample the possible values K times. For each sampled value $x_{i,k}$, we adopt MCS to sample the missing values and compute the uncertainty with the standard deviation of the outputs (denoted as $U_{i,k}$). If the variable i is observed, the new uncertainty would be $U_i = \frac{1}{K}\sum_{k=1}^{K} U_{i,k}$. We select the variables with the maximal uncertainty reduction: $i^* = \arg\max_i U - U_i = \arg\min_i U_i$. We set $k = 500$ in our implementation. The expected uncertainty decrease is $U - U_i$ for variable i. With such a model design, the front-end UI can support interactive visualization that allows users to explore the different laboratory test's effectiveness and further satisfies **Design Strategy 5**.

4.4 System Implementation

The interactive front-end user interface (Section 4.2) is developed as a web application using the React framework. The particular visualization that enables users to interactively explore the laboratory test recommendations (Section 4.3) is developed using the. Recharts [2] library. We developed the back-end with Python's FastAPI [3] library. The predictive model and the counterfactual model inside the back-end (Section 4.3.1) are implemented with PyTorch [61]. We primarily store the data (i.e., MIMIC-III [37]) in a master-slave backup MySQL database for efficient querying and security purposes. The entire prototype system (front-end, back-end, model, and database) is hosted on Amazon Web Service (AWS) server instances.

5 EVALUATION STUDY: ENHANCING HUMAN-AI COLLABORATION

Our system aims to implement a human-AI collaboration paradigm for the decision-making scenario for sepsis diagnosis. To do so, we design our SepsisLab system that shifts the focus of the AI predictions to offering human experts their desired and actionable recommendations in the intermediate stages of a medical decision-making workflow. In this section, we report on a heuristic evaluation study by inviting the same six clinicians to interact with and provide feedback on our SepsisLab design. Our findings demonstrate that SepsisLab is generally appreciated as it can provide meaningful assistance to clinicians. The participants would love to see such a system deployed into the real EPIC EHR system, and they argue that our AI design can help them achieve a better human-AI team experience.

5.1 Design and Procedure

We recruited the same six clinical experts in Section 3 to perform a heuristic test with our system. We first pre-loaded de-identified patient data from MIMIC-III [37] into our prototype and deployed it in an internal cloud cluster to prevent leakage of patient data. To demonstrate the system across varied patient conditions, we selected two patient groups from MIMIC-III [37], one with patients ultimately diagnosed with sepsis and the other without sepsis. Patients ultimately diagnosed with sepsis were selected based on the sepsis-3 [75] criteria and are all adults. Limited by the duration of each study session, we randomly selected five data points from each patient group, totaling ten data points for display and interaction with participants. Due to the usage regulations that researchers need approvals to access for MIMIC-III [37], we only present the mock data in the paper to present the UI. Our SepsisLab system was used as a design probe to solicit participants' user experience and design requirements. Participants navigated the interface and thought aloud in testing its different functionalities. We then conducted a semi-structured post-study interview. Specifically, we asked about their comments on the AI assistance's new focus on intermediate decision-making steps, and whether the lab test recommendations and the counterfactual predictions were practical and could enhance their decision-making process. We also collected clinicians' suggestions for improvements to the system. Each user study session lasted around 30 minutes. We recorded the interview audio and used an inductive process [82] to analyze the interview transcription.

5.2 Result

Overall, the participants gave very positive comments on our system prototype. With AI shifting focus from final-stage prediction to intermediate stages, clinicians found AI less intimidating but more

[2] https://recharts.org/
[3] https://fastapi.tiangolo.com/

cooperative. Meanwhile, participants also liked our new functions that went beyond risk score prediction and commented that they were helpful in revealing more information, providing actionable insights and improving AI transparency.

5.2.1 Shifting AI Focus Away from Final Decision Prediction Can Enhance Human-AI Collaboration Experience. As revealed from our formative study, the fundamental challenge of the existing sepsis module lies in the fact that it only focuses on the final stage of the four-step diagnosis process (see Section 3.2.5). Our system aims to address this challenge by shifting AI's focus from the last stage to the intermediate stages (*hypotheses generation and data gathering for hypothesis testing*). These are the steps where physicians need more assistance. Our evaluation study results confirmed that the new system was able to improve the human-AI collaboration experience.

Clinicians (P2, P6) commented that the early prediction with uncertainty visualization and lab test recommendations could ease feelings of being in competition with or replaced by AI, since it no longer focuses on the final stage and leaves humans to decide whether to take the suggestions or not. Instead, our system ensures that physicians retain the role of final decision-maker.

> *"if you tell me, you (model) need these labs, you got it, buddy, I will order those labs. It's not a problem..."* (P2)

Our design probe significantly alleviated experts' concerns about the threat of AI. Participants felt that our new sepsis module offers the experience of teaming with AI rather than competition, creating an actual human-AI collaboration paradigm.

> *"I think knowing what [lab results] would help AI to have a better prediction would let me give it a try, and [the lab results] may be more useful in my decision."* (P6)

The shift of AI assistance's focus and the improvement in the collaboration experience is combining efforts of multiple functions supported by our system. In the rest of this section, we summarize participants' feedback on each function of our system.

5.2.2 Future Prediction and Visualization Reveals More Information. All of the six participants in our study liked the time-based prediction capability of our system. The prediction graph, with the uncertainty visualization, went beyond the final decision stage and provided much richer information than a risk score *"I love that graph because it says the whole picture."* (P1) This could help physicians better understand the model and inspect whether it is reliable. Participants agreed that adding more explanations to charts would help them make better decisions.

> *"So I think having some of that [temporal] information is helpful to understand what drives the model and also where it's gonna go in 10 hours from now, if that model is good and consistent... I would hope that somebody could show that this where the model is gonna go [in the EPIC sepsis system], as it matches up with what happens in real life. "* (P5)

On the one hand, this is supported by previous work about the advantages and benefits of XAI [87, 105]. On the other hand, the future prediction function embeds the shift of AI focus from final diagnosis results to intermediate suggestions (help experts propose and test hypotheses). This leaves enough flexible space for human experts to make intermediate and final decisions. Meanwhile, the system also places more emphasis on uncertainty estimation. The participants agreed that providing these aspects of information was much more helpful than a single risk score.

> *"Yeah, I think that will be much better than just giving me just a score. If I can see something like this [counterfactual explanation], so the nurse gets this, or I am opening it up, I'm seeing that, okay, this patient came in with a risk score of five and now [with the suggested lab result] it is at 15. And based on the model, in the next eight hours, or 10 hours, it's going to be like 25 or 30. So, that is more meaningful information than just a score."* (P3)

5.2.3 Laboratory Test Recommendation Provides Actionable Insights. All six participants responded positively to the lab test recommendation. First of all, participants confirmed that the lab test recommendations were consistent with current physician workflows. This validates the design of this system function. Meanwhile, participants unanimously said that they would take advantage of this recommendation function to help them consider and act on laboratory tests.

> *"If I was able to click on the more lab tests needed and get an idea on how much it could narrow. And if it was something where it recommended these diagnostics, and I did these diagnostics, and when the labs came back, somehow triggered me to look at it again... and that prediction [uncertainty] has narrowed, [prediction] got very accurate in terms of risk score... This would be cool."* (P2)

Combined with the future prediction function, this could provide clinicians with more insights than a risk score or alert, helping them narrow down their hypotheses and make better sepsis diagnostic decisions.

> *"In the ICU, that [lab recommendation] would be more realistic as a flag than the sepsis alert that we have right now, because it's telling you something. It's giving you information."* (P1)

These results suggest that our system design has the potential to provide actionable insights to health experts and support the stages from generating hypotheses to gathering data to testing hypotheses.

5.2.4 Counterfactual Information Improves Transparency. Closely related to the laboratory test recommendation function, the participants agreed that showing counterfactual predictions based on recommendation would drive their choice of laboratory tests. Traditionally, reliance on clinicians' personal experience was common, yet diagnosing sepsis posed significant challenges due to its inherent uncertainty. With our system prototype, they commented that being able to show potential outcomes of different lab tests would help them with the thinking process and narrow down the search space.

> *If the system is going to alert me: hey, maybe you should repeat a lactate now because your patient is scoring high;*

> *Or maybe you should repeat a white blood count today because you don't have lab work today; or something like that. That will certainly be helpful.* (P3)

Moreover, the participants (P2, P4) also mentioned that showing counterfactual information also helped them better understand the model's functions and its future prediction process.

> *"And then again, having it [the alternative trajectory with a counterfactual lab result] helps me with what AI wants me to order, so that it can be more accurate, [this design] is good."* (P2)

> *"Especially if you're asking me to order a repeat lactate, telling me that, hey, your sepsis score was a five because your lactate was high and your white count was high. And I [AI] want that lactate, and [ordering a lactate] is going to help me kind of closing down my model. I think that'll be useful. "* (P4)

5.2.5 Areas for Future Improvement. In addition to the positive comments through the design probe with our system, participants also suggested a few concerns and expectations about our system for future real-world deployment. One concern is the performance of AI predictions and recommendations. With our algorithm having a stronger capability than a risk score, the model becomes more complex, and participants were concerned about its reliability. This is commonly observed in many health-focused AI systems [5, 86].

Meanwhile, participant (P3) also mentioned the risk of information overload. *"But simultaneously, you have to be careful and mindful of giving too much information at once. So maybe it can be a step-wise approach."* (P3) This suggests a future improvement direction of our system to that providing information step by step that follows clinicians through their decision-making process and provides appropriate explanations will best assist them in making a diagnosis.

In addition, some participants (P4, P5) also mentioned the potential to expand the usage scenarios of the counterfactual predictions and explanations, such as showing the potential impact of certain treatments on the development of a patient's symptom conditions. Although this is beyond the scope of this paper, we discuss a few promising directions of future work in Section 6.

6 DISCUSSION

In this section, we discuss the design implications obtained from our interview and evaluation studies. We then discuss the new human-AI collaboration paradigm and its application beyond sepsis diagnosis. We also highlight the risks and ethical concerns associated with the paradigm, as well as the limitations of our work.

6.1 Human-AI Collaboration in High-Uncertainty, High-Stakes, and Time-Sensitive Decision Making

With sepsis diagnosis as an example, our work reveals a fundamental problem in the existing human-AI collaboration paradigm for high-uncertainty, high-stakes, and time-sensitive decision-making tasks. Most AI research in this space aims to create the most accurate risk score prediction model. However, this goal is unable to form an effective human-AI collaboration. Our study shows that physicians found such a risk prediction model unhelpful and intimidating, challenging their role as the final decision-maker.

Instead, we argue that AI should aim to support human experts in the intermediate stages (Step 1 - 3 in Figure 1) rather than the final stage. This can position AI in an appropriate place to effectively support experts' decision-making process, while not influencing their decision-maker role. In our case, we introduce a novel approach by providing future prediction and uncertainty visualization, lab test recommendations, and counterfactual information and support (rather than challenge) experts' final decisions. This establishes a unique form of 'communication and collaboration' between physicians and the AI by providing actionable recommendations. Our evaluation study results suggest that such a shift in AI's focus can indeed create a new human-AI team paradigm. Our exploration into sepsis diagnostic decision-making exemplifies this collaborative process. In practice, this does not conflict with current XAI research and provides a new perspective on the role that AI can play in the decision-making process.

6.2 Beyond Sepsis Diagnosis

We envision such a new human-AI collaboration paradigm can move beyond sepsis diagnosis. In healthcare, there are a number of decision-making tasks that have similar properties as sepsis diagnosis: highly uncertain, high-stakes, and time-sensitive [41]. Examples include both physical health problems (e.g., stroke, heart attack, and meningitis), and mental health problems (e.g., major depressive disorder with suicidal ideation, bipolar disorder during a manic episode, and schizophrenia with psychosis). In these cases, the symptoms tend to be ambiguous, noisy, and highly individual, while having life-threatening consequences if not treated in a timely manner. If an AI only focuses on predicting the outcome of the final decision stage, expert clinicians would also find it as a "challenging and intimidating competitor" rather than a collaborator and a partner, leading to the abortion of AI. Our design can potentially be generalized to these fields, where AI should also support these experts in their intermediate decision-making stages and help them propose hypotheses, gather information, and test hypotheses.

In addition, we foresee that such a new human-AI collaboration paradigm can be applicable to other non-healthcare complex decision-making scenarios as well, such as military (e.g., hostage rescues, evacuation operations), business (e.g., product launch/recalls, market crisis), emergency response (e.g., earthquake response, wildfire response), just name a few. All these cases require fast and accurate human decisions to reduce uncertainty and achieve optimal outcomes. Our proposal of the new human-AI collaboration paradigm can inspire the existing solutions in these fields to shift their AI focus to better support domain experts.

6.3 Risks and Ethical Concerns of AI-powered Decision Making

Despite the promising advantage of our newly proposed human-AI collaboration paradigm, we also want to highlight the risks and ethical concerns associated with it. For example, such a new paradigm would introduce additional burden to experts [13, 56, 72]. In our study, participants were concerned about the cognitive load caused by our system, which has been reported in previous studies

related to visualization [2, 36, 86]. Meanwhile, mistakes and errors made by AI are inevitable, and the potential biases embedded in AI algorithms are not yet addressed by this new paradigm. The responsibility still falls on human experts to minimize these risks and biases through rigorous testing and evaluation. Besides, there is a potential risk of over-dependence on AI systems, which might foster complacency and reduce vigilance among human experts. This is an open research question for future researchers, requiring a balanced design approach that promotes collaboration while avoiding an undue reliance on AI recommendations.

6.4 Limitation and Future Work

There are several limitations in our work. First, our study population is limited. We only involved six physicians in our formative study and heuristic evaluation (similar to the number of expert participants in prior work's formative study [10]), who came from the same hospital and only used one specific sepsis module ESM. Although it is one of the mostly commonly used sepsis modules in the U.S., there could be some systematic biases in our study results. Future work needs to involve more diverse populations from multiple hospitals. Second, our system is implemented as a prototype and not integrated into the EHR system. Our evaluation study used our system as a design probe to collect clinicians' feedback. This may influence the validity and generalizability of our results. Our findings may be different if our system is actually deployed in the real world. We picked ESM as a case study, since it has been the focus of extensive prior research as an early detection system [19]. Further research is required to assess additional sepsis early detection systems and comparable tools for early disease identification to enhance understanding of the clinical decision-making process. Moreover, to further elucidate the challenges arising from AI-assisted decision-making and to develop systems that are more congruent with the clinical decision-making processes, quantitative research and additional clinical practices are required in the future. Third, our algorithm and visualization have room for improvement. As mentioned in Section 4, there are more visualization methods and algorithms for risk prediction and uncertainty estimation. Future work can explore the effectiveness of more back-end techniques.

Furthermore, in our interview, the participants mentioned that they barely relied on the existing sepsis module ESM in their decision-making process. Their comments reveal that this module needs a better design to support clinicians' workflow. Our prototype in Section 4 presents an initial step towards a better design. And there are a few more directions to improve. Based on the comments mentioned by the participants in Section 5.2.5, a future sepsis module should provide a simple and easy-to-operate interface. Clinicians are usually over-loaded [36, 86], and an appropriately designed interface could improve their efficiency.

We also find that physicians and nurses often have different responsibilities in the diagnostic decision-making process. In our case, nurses are tasked with receiving alerts and determining if they warrant escalation to physicians, while physicians make the ultimate decision on the necessary subsequent actions. Different workflows that arise from different clinical roles are often overlooked in contemporary AI-CDSS designs. However, predominant EHR systems and AI-CDSS platforms fail to distinguish between these distinct roles and tend to offer a one-size-fits-all interface and set of functionalities to both physicians and nurses. In future designs, the systems should be role-specific and tailored to both physicians and nurses, ensuring that each can extract relevant information from model predictions. This not only enhances the system's efficiency but also ensures that each medical professional is equipped with the right tools to make informed decisions.

7 CONCLUSION

In this work, we aim to design a better human-AI collaboration paradigm to support human experts in high-uncertainty, high-stakes, and time-sensitive decision-making tasks. We focus on sepsis diagnosis, a common yet life-threatening disease. We conducted a formative study with six physicians with rich sepsis-treating experience to better understand the existing challenges of human-AI collaboration with a common sepsis module. Our results reveal that the existing module is not only useless but also leads to additional meaningless workloads. More importantly, it reveals a fundamental problem of the wrong focus of AI: AI should not focus on predicting or suggesting the final decision-making stage, which could be challenging and intimidating to human experts as the final decision-maker. Based on these insights, we developed a system that aims to address these challenges. Our new system, SepsisLab, shifts the AI's focus from the final stage to the intermediate stages (generating hypotheses, gathering information, and testing hypotheses). SepsisLab improves a sepsis diagnosis algorithm with future prediction and uncertainty visualization, provides lab test recommendations, and offers counterfactual information. Our evaluation study shows that the new system prototype can provide actionable insights, improve transparency, and better support the clinicians' decision-making process, forming a new human-AI collaboration paradigm. We envision that our findings can shed light on the design of better human-AI collaboration paradigms for other scenarios with complex decision-making tasks.

ACKNOWLEDGMENTS

This work was funded in part by the National Science Foundation under award number IIS-2145625 and by the National Institutes of Health under award number R01GM141279 and R01MD018424. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation or the National Institutes of Health. The authors extend heartfelt thanks to the participants from OSU Wexner Medical Center.

REFERENCES

[1] Julia Amann, Dennis Vetter, Stig Nikolaj Blomberg, Helle Collatz Christensen, Megan Coffee, Sara Gerke, Thomas K Gilbert, Thilo Hagendorff, Sune Holm, Michelle Livne, et al. 2022. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. *PLOS Digital Health* 1, 2 (2022), e0000016.

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300233

[3] Inci M Baytas, Cao Xiao, Xi Zhang, Fei Wang, Anil K Jain, and Jiayu Zhou. 2017. Patient subtyping via time-aware LSTM networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 65–74.

[4] Armando D Bedoya, Meredith E Clement, Matthew Phelan, Rebecca C Steorts, Cara O'Brien, and Benjamin A Goldstein. 2019. Minimal impact of implemented early warning score and best practice alert for patient deterioration. *Critical care medicine* 47, 1 (2019), 49.

[5] 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. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[6] Roger C Bone, Robert A Balk, Frank B Cerra, R Phillip Dellinger, Alan M Fein, William A Knaus, Roland MH Schein, and William J Sibbald. 1992. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest* 101, 6 (1992), 1644–1655.

[7] Eleanor R. Burgess, Ivana Jankovic, Melissa Austin, Nancy Cai, Adela Kapuścińska, Suzanne Currie, J. Marc Overhage, Erika S Poole, and Jofish Kaye. 2023. Healthcare AI Treatment Decision Support: Design Principles to Enhance Clinician Adoption and Trust. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 15, 19 pages. https://doi.org/10.1145/3544548.3581251

[8] Estefanía Caballero-Ruiz, Gema García-Sáez, Mercedes Rigla, María Villaplana, Belen Pons, and M Elena Hernando. 2017. A web-based clinical decision support system for gestational diabetes: Automatic diet prescription and detection of insulin needs. *International journal of medical informatics* 102 (2017), 35–49.

[9] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–14.

[10] 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. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 104 (nov 2019), 24 pages. https://doi.org/10.1145/3359206

[11] Maurizio Cecconi, Laura Evans, Mitchell Levy, and Andrew Rhodes. 2018. Sepsis and septic shock. *The Lancet* 392, 10141 (2018), 75–87.

[12] Ahmad Chaddad, Jihao Peng, Jian Xu, and Ahmed Bouridane. 2023. Survey of explainable AI techniques in healthcare. *Sensors* 23, 2 (2023), 634.

[13] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F. Maxwell Harper, and Haiyi Zhu. 2019. Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300789

[14] Yu Cheng, Fei Wang, Ping Zhang, and Jianying Hu. 2016. Risk prediction with electronic health records: A deep learning approach. In *Proceedings of the 2016 SIAM international conference on data mining*. SIAM, 432–440.

[15] Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. 2023. Are Two Heads Better Than One in AI-Assisted Decision Making? Comparing the Behavior and Performance of Groups and Individuals in Human-AI Collaborative Recidivism Risk Assessment. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 348, 18 pages. https://doi.org/10.1145/3544548.3581015

[16] Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. GRAM: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 787–795.

[17] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. 2016. Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. *Advances in neural information processing systems* 29 (2016).

[18] Limeng Cui, Haeseung Seo, Maryam Tabar, Fenglong Ma, Suhang Wang, and Dongwon Lee. 2020. DETERRENT: Knowledge Guided Graph Attention Network for Detecting Healthcare Misinformation. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 492–502. https://doi.org/10.1145/3394486.3403092

[19] John Cull, Robert Brevetta, Jeff Gerac, Shanu Kothari, and Dawn Blackhurst. 2023. Epic Sepsis Model Inpatient Predictive Analytic Tool: A Validation Study. *Critical Care Explorations* 5, 7 (2023).

[20] Alexa Dierig, Christoph Berger, Philipp K. A. Agyeman, Sara Bernhard-Stirnemann, Eric Giannoni, Martin Stocker, Klara M. Posfay-Barbe, Anita Niederer-Loher, Christian R. Kahlert, Alex Donas, Paul Hasters, Christa Relly, Thomas Riedel, Christoph Aebi, Luregn J. Schlapbach, and Swiss Pediatric Sepsis Study Heininger, Ulrich and. 2018. Time-to-Positivity of Blood Cultures in Children With Sepsis. *Frontiers in Pediatrics* 6 (2018). https://doi.org/10.3389/fped.2018.00222

[21] Maia Dorsett, Melissa Kroll, Clark S Smith, Phillip Asaro, Stephen Y Liang, and Hawnwan P Moy. 2017. qSOFA has poor sensitivity for prehospital identification of severe sepsis and septic shock. *Prehospital emergency care* 21, 4 (2017), 489–497.

[22] Norman Lance Downing, Joshua Rolnick, Sarah F Poole, Evan Hall, Alexander J Wessels, Paul Heidenreich, and Lisa Shieh. 2019. Electronic health record-based clinical decision support alert for severe sepsis: a randomised evaluation. *BMJ quality & safety* 28, 9 (2019), 762–768.

[23] Yongping Du, Jingya Yan, Yuxuan Lu, Yiliang Zhao, and Xingnan Jin. 2023. Improving Biomedical Question Answering by Data Augmentation and Model Weighting. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 20, 2 (2023), 1114–1124. https://doi.org/10.1109/TCBB.2022.3171388

[24] Juan Manuel Durán and Karin Rolanda Jongsma. 2021. Who is afraid of black box algorithms? On the epistemological and ethical basis of trust in medical AI. *Journal of Medical Ethics* 47, 5 (2021), 329–335. https://doi.org/10.1136/medethics-2020-106820 arXiv:https://jme.bmj.com/content/47/5/329.full.pdf

[25] Upol Ehsan, Philipp Wintersberger, Q. Vera Liao, Elizabeth Anne Watkins, Carina Manger, Hal Daumé III, Andreas Riener, and Mark O Riedl. 2022. Human-Centered Explainable AI (HCXAI): Beyond Opening the Black-Box of AI. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI EA '22)*. Association for Computing Machinery, New York, NY, USA, Article 109, 7 pages. https://doi.org/10.1145/3491101.3503727

[26] Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M Coopersmith, Craig French, Flávia R Machado, Lauralyn Mcintyre, Marlies Ostermann, Hallie C Prescott, et al. 2021. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock 2021. *Intensive care medicine* 47, 11 (2021), 1181–1247.

[27] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1362–1374. https://doi.org/10.1145/3531146.3533193

[28] 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.

[29] Kim Huat Goh, Le Wang, Adrian Yong Kwang Yeow, Hermione Poh, Ke Li, Joannas Jie Lin Yeow, and Gamaliel Yu Heng Tan. 2021. Artificial intelligence in sepsis early prediction and diagnosis using unstructured data in healthcare. *Nature communications* 12, 1 (2021), 711.

[30] Leo A Goodman. 1961. Snowball sampling. *The annals of mathematical statistics* (1961), 148–170.

[31] Guillermo Gutierrez. 2020. Artificial intelligence in the intensive care unit. *Annual Update in Intensive Care and Emergency Medicine 2020* (2020), 667–681.

[32] Anand R. Habib, Anthony L. Lin, and Richard W. Grant. 2021. The Epic Sepsis Model Falls Short—The Importance of External Validation. *JAMA Internal Medicine* 181, 8 (08 2021), 1040–1041. https://doi.org/10.1001/jamainternmed.2021.3333

[33] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[34] Pulwasha Iftikhar, Marcela V Kuijpers, Azadeh Khayyat, Aqsa Iftikhar, and Maribel DeGouvia De Sa. 2020. Artificial intelligence: a new paradigm in obstetrics and gynecology research and clinical practice. *Cureus* 12, 2 (2020).

[35] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C. Ahn, Thomas H. McCoy, Roy H. Perlis, Finale Doshi-Velez, and Krzysztof Z. Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 659, 14 pages. https://doi.org/10.1145/3411764.3445385

[36] Zhuochen Jin, Shuyuan Cui, Shunan Guo, David Gotz, Jimeng Sun, and Nan Cao. 2020. CarePre: An Intelligent Clinical Decision Assistance System. *ACM Trans. Comput. Healthcare* 1, 1, Article 6 (mar 2020), 20 pages. https://doi.org/10.1145/3344258

[37] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data* 3, 1 (2016), 1–9.

[38] Krishna Juluru, Hao-Hsin Shih, Krishna Nand Keshava Murthy, Pierre Elnajjar, Amin El-Rowmeim, Christopher Roth, Brad Genereaux, Josef Fox, Eliot Siegel, and Daniel L Rubin. 2021. Integrating Al algorithms into the clinical workflow. *Radiology: Artificial Intelligence* 3, 6 (2021), e210013.

[39] Sundreen Asad Kamal, Changchang Yin, Buyue Qian, and Ping Zhang. 2020. An interpretable risk prediction model for healthcare with pattern attention. *BMC Medical Informatics and Decision Making* 20 (2020), 1–10.

[40] Hwan Il Kim and Sunghoon Park. 2019. Sepsis: early recognition and optimized treatment. *Tuberculosis and respiratory diseases* 82, 1 (2019), 6–14.

[41] 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. *arXiv preprint arXiv:2112.11471* (2021).

[42] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. https://doi.org/10.1145/3287560.3287590

[43] Min Hun Lee, Daniel P. Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez Bermúdez i Badia. 2021. A Human-AI Collaborative Approach for Clinical Decision Making on Rehabilitation Assessment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 392, 14 pages. https://doi.org/10.1145/3411764.3445472

[44] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684. https://doi.org/10.1177/2053951718756684 arXiv:https://doi.org/10.1177/2053951718756684

[45] Min Kyung Lee and Katherine Rich. 2021. Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 138, 14 pages. https://doi.org/10.1145/3411764.3445570

[46] Manuel Lentzen, Sumit Madan, Vanessa Lage-Rupprecht, Lisa Kühnel, Juliane Fluck, Marc Jacobs, Mirja Mittermaier, Martin Witzenrath, Peter Brunecker, Martin Hofmann-Apitius, et al. 2022. Critical assessment of transformer-based AI models for German clinical notes. *JAMIA open* 5, 4 (2022), ooac087.

[47] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. https://doi.org/10.1145/3313831.3376590

[48] Chen Lin, Yuan Zhang, Julie Ivy, Muge Capan, Ryan Arnold, Jeanne M Huddleston, and Min Chi. 2018. Early diagnosis and prediction of sepsis shock by combining static and dynamic information using convolutional-LSTM. In *2018 IEEE international conference on healthcare informatics (ICHI)*. IEEE, 219–228.

[49] R. Liu, J.L. Greenstein, S.J. Granite, J.C. Fackler, M.M. Bembea, S.V. Sarma, and R.L. Winslow. 2019. Data-driven discovery of a novel sepsis pre-shock state predicts impending septic shock in the ICU. *Sci. Rep.* 9, 1 (2019), 1–9. https://doi.org/10.1038/s41598-019-42637-5

[50] Robyn Longhurst. 2003. Semi-structured interviews and focus groups. *Key methods in geography* 3, 2 (2003), 143–156.

[51] Yuxuan Lu, Jingya Yan, Zhixuan Qi, Zhongzheng Ge, and Yongping Du. 2022. Contextual Embedding and Model Weighting by Fusing Domain Knowledge on Biomedical Question Answering. In *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (Northbrook, Illinois) *(BCB '22)*. Association for Computing Machinery, New York, NY, USA, Article 54, 4 pages. https://doi.org/10.1145/3535508.3545508

[52] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 78, 16 pages. https://doi.org/10.1145/3411764.3445562

[53] Patrick G Lyons, Mackenzie R Hofford, C Yu Sean, Andrew P Michelson, Philip RO Payne, Catherine L Hough, and Karandeep Singh. 2023. Factors associated with variability in the performance of a proprietary sepsis prediction model across 9 networked hospitals in the US. *JAMA Internal Medicine* (2023).

[54] Fenglong Ma, Radha Chitta, Jing Zhou, Quanzeng You, Tong Sun, and Jing Gao. 2017. Dipole: Diagnosis prediction in healthcare via attention-based bidirectional recurrent neural networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*. 1903–1911.

[55] Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. Kame: Knowledge-based attention model for diagnosis prediction in healthcare. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 743–752.

[56] Swati Mishra and Jeffrey M Rzeszotarski. 2021. Designing Interactive Transfer Learning Tools for ML Non-Experts. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 364, 15 pages. https://doi.org/10.1145/3411764.3445096

[57] Su Q Nguyen, Edwin Mwakalindile, James S Booth, Vicki Hogan, Jordan Morgan, Charles T Prickett, John P Donnelly, and Henry E Wang. 2014. Automated electronic medical record sepsis detection in the emergency department. *PeerJ* 2 (2014), e343.

[58] Yuka Otaki, Ananya Singh, Paul Kavanagh, Robert JH Miller, Tejas Parekh, Balaji K Tamarappoo, Tali Sharir, Andrew J Einstein, Mathews B Fish, Terrence D Ruddy, et al. 2022. Clinical deployment of explainable artificial intelligence of SPECT for diagnosis of coronary artery disease. *Cardiovascular Imaging* 15, 6 (2022), 1091–1102.

[59] Lace Padilla, Racquel Fygenson, Spencer C Castro, and Enrico Bertini. 2022. Multiple forecast visualizations (mfvs): Trade-offs in trust and performance in multiple covid-19 forecast visualizations. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2022), 12–22.

[60] Lace Padilla, Matthew Kay, and Jessica Hullman. 2021. *Uncertainty Visualization*. John Wiley & Sons, Ltd, 1–18. https://doi.org/10.1002/9781118445112.stat08296 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118445112.stat08296

[61] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.), Vol. 32. Curran Associates, Inc.

[62] W Nicholson Price. 2018. Big data and black-box medical algorithms. *Science translational medicine* 10, 471 (2018), eaao5333.

[63] Alireza Rafiei, Alireza Rezaee, Farshid Hajati, Soheila Gheisari, and Mojtaba Golzan. 2021. SSP: Early prediction of sepsis using fully connected LSTM-CNN model. *Computers in biology and medicine* 128 (2021), 104110.

[64] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. 2019. Machine learning in medicine. *New England Journal of Medicine* 380, 14 (2019), 1347–1358.

[65] Santiago Romero-Brufau, Kirk D Wyatt, Patricia Boyum, Mindy Mickelson, Matthew Moore, and Cheristi Cognetta-Rieke. 2020. A lesson in implementation: a pre-post study of providers' experience with artificial intelligence-based clinical decision support. *International journal of medical informatics* 137 (2020), 104072.

[66] Kristina E Rudd, Sarah Charlotte Johnson, Kareha M Agesa, Katya Anne Shackelford, Derrick Tsoi, Daniel Rhodes Kievlan, Danny V Colombara, Kevin S Ikuta, Niranjan Kissoon, Simon Finfer, et al. 2020. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the Global Burden of Disease Study. *The Lancet* 395, 10219 (2020), 200–211.

[67] Mohammed Saqib, Ying Sha, and May D Wang. 2018. Early prediction of sepsis in EMR records using traditional ML techniques and deep learning LSTM networks. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 4038–4041.

[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. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 99–109.

[69] 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. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 99–109. https://doi.org/10.1145/3351095.3372827

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

[71] Claude E Shannon and Warren Weaver. 1949. A mathematical model of communication. *Urbana, IL: University of Illinois Press* 11 (1949), 11–20.

[72] Hong Shen, Haojian Jin, Ángel Alexander Cabrera, Adam Perer, Haiyi Zhu, and Jason I. Hong. 2020. Designing Alternative Representations of Confusion Matrices to Support Non-Expert Public Understanding of Algorithm Performance. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 153 (oct 2020), 22 pages. https://doi.org/10.1145/3415224

[73] Ben Shneiderman and Catherine Plaisant. 2010. *Designing the user interface: Strategies for effective human-computer interaction*. Pearson Education India.

[74] Jasmine A Silvestri, Tyler E Kmiec, Nicholas S Bishop, Susan H Regli, and Gary E Weissman. 2022. Desired Characteristics of a Clinical Decision Support System for Early Sepsis Recognition: Interview Study Among Hospital-Based Clinicians. *JMIR Hum Factors* 9, 4 (21 Oct 2022), e36976. https://doi.org/10.2196/36976

[75] Mervyn Singer, Clifford S. Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R. Bernard, Jean-Daniel Chiche, Craig M. Coopersmith, Richard S. Hotchkiss, Mitchell M. Levy, John C. Marshall, Greg S. Martin, Steven M. Opal, Gordon D. Rubenfeld, Tom van der Poll, Jean-Louis Vincent, and Derek C. Angus. 2016. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA* 315, 8 (02 2016), 801–810. https://doi.org/10.1001/jama.2016.0287 arXiv:https://jamanetwork.com/journals/jama/articlepdf/2492881/jsc160002.pdf

[76] 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* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 754, 18 pages. https://doi.org/10.1145/3544548.3581075

[77] Gary B Smith, David R Prytherch, Paul Meredith, Paul E Schmidt, and Peter I Featherstone. 2013. The ability of the National Early Warning Score (NEWS) to discriminate patients at risk of early cardiac arrest, unanticipated intensive care unit admission, and death. *Resuscitation* 84, 4 (2013), 465–470.

[78] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3313831.3376624

[79] Harold C Sox, Michael C. Higgins, and Douglas K. Owens. 2013. *Medical Decision Making*. John Wiley & Sons, Ltd.

[80] Christian P Subbe, Michael Kruger, Peter Rutherford, and L Gemmel. 2001. Validation of a modified Early Warning Score in medical admissions. *Qjm* 94, 10 (2001), 521–526.

[81] Myriam Tanguay-Sela, David Benrimoh, Christina Popescu, Tamara Perez, Colleen Rollins, Emily Snook, Eryn Lundrigan, Caitrin Armstrong, Kelly Perlman, Robert Fratila, et al. 2022. Evaluating the perceived utility of an artificial intelligence-powered clinical decision support system for depression treatment using a simulation center. *Psychiatry Research* 308 (2022), 114336.

[82] David R Thomas. 2006. A general inductive approach for analyzing qualitative evaluation data. *American journal of evaluation* 27, 2 (2006), 237–246.

[83] Daniel SW Ting, Yong Liu, Philippe Burlina, Xinxing Xu, Neil M Bressler, and Tien Y Wong. 2018. AI for medical imaging goes deep. *Nature medicine* 24, 5 (2018), 539–540.

[84] Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* 25, 1 (2019), 44–56.

[85] Omar A Usman, Asad A Usman, and Michael A Ward. 2019. Comparison of SIRS, qSOFA, and NEWS for the early identification of sepsis in the Emergency Department. *The American journal of emergency medicine* 37, 8 (2019), 1490–1497.

[86] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 697, 18 pages. https://doi.org/10.1145/3411764.3445432

[87] Xinru Wang and Ming Yin. 2021. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In *26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 318–328. https://doi.org/10.1145/3397481.3450650

[88] Andrew Wong, Erkin Otles, John P Donnelly, Andrew Krumm, Jeffrey McCullough, Olivia DeTroyer-Cooley, Justin Pestrue, Marie Phillips, Judy Konye, Carleen Penoza, et al. 2021. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. *JAMA Internal Medicine* 181, 8 (2021), 1065–1070.

[89] World Health Organization. 2020. Global report on the epidemiology and burden of sepsis: current evidence, identifying gaps and future directions. (2020).

[90] Wei Xiong, Hongmiao Fan, Liang Ma, and Chen Wang. 2022. Challenges of human—machine collaboration in risky decision-making. *Frontiers of Engineering Management* 9, 1 (2022), 89–103.

[91] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K. Villalba, Janine M. Dutcher, Michael J. Tumminia, Tim Althoff, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Jennifer Mankoff, and Anind K. Dey. 2019. Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 116 (sep 2019), 33 pages. https://doi.org/10.1145/3351274

[92] Xuhai Xu, Xin Liu, Han Zhang, Weichen Wang, Subigya Nepal, Yasaman Sefidgar, Woosuk Seo, Kevin S. Kuehn, Jeremy F. Huckins, Margaret E. Morris, Paula S. Nurius, Eve A. Riskin, Shwetak Patel, Tim Althoff, Andrew Campbell, Anind K. Dey, and Jennifer Mankoff. 2023. GLOBEM: Cross-Dataset Generalization of Longitudinal Human Behavior Modeling. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 4, Article 190 (jan 2023), 34 pages. https://doi.org/10.1145/3569485

[93] Xuhai Xu, Han Zhang, Yasaman Sefidgar, Yiyi Ren, Xin Liu, Woosuk Seo, Jennifer Brown, Kevin Kuehn, Mike Merrill, Paula Nurius, Shwetak Patel, Tim Althoff, Margaret Morris, Eve Riskin, Jennifer Mankoff, and Anind Dey. 2022. GLOBEM Dataset: Multi-Year Datasets for Longitudinal Human Behavior Modeling Generalization. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 24655–24692.

[94] Qian Yang, Yuexing Hao, Kexin Quan, Stephen Yang, Yiran Zhao, Volodymyr Kuleshov, and Fei Wang. 2023. Harnessing Biomedical Literature to Calibrate Clinicians' Trust in AI Decision Support Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 14, 14 pages. https://doi.org/10.1145/3544548.3581393

[95] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–11.

[96] Qian Yang, John Zimmerman, Aaron Steinfeld, Lisa Carey, and James F Antaki. 2016. Investigating the heart pump implant decision process: opportunities for decision support tools to help. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4477–4488.

[97] Changchang Yin, Ruoqi Liu, Jeffrey Caterino, and Ping Zhang. 2022. Deconfounding Actor-Critic Network with Policy Adaptation for Dynamic Treatment Regimes. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Washington DC, USA) *(KDD '22)*. Association for Computing Machinery, New York, NY, USA, 2316–2326. https://doi.org/10.1145/3534678.3539413

[98] Changchang Yin, Ruoqi Liu, Dongdong Zhang, and Ping Zhang. 2020. Identifying Sepsis Subphenotypes via Time-Aware Multi-Modal Auto-Encoder. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (Virtual Event, CA, USA) *(KDD '20)*. Association for Computing Machinery, New York, NY, USA, 862–872. https://doi.org/10.1145/3394486.3403129

[99] Changchang Yin, Rongjian Zhao, Buyue Qian, Xin Lv, and Ping Zhang. 2019. Domain knowledge guided deep learning with electronic health records. In *2019 IEEE International Conference on Data Mining (ICDM)*. IEEE, 738–747.

[100] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. 2018. Artificial intelligence in healthcare. *Nature biomedical engineering* 2, 10 (2018), 719–731.

[101] Kuo-Ching Yuan, Lung-Wen Tsai, Ko-Han Lee, Yi-Wei Cheng, Shou-Chieh Hsu, Yu-Sheng Lo, and Ray-Jade Chen. 2020. The development an artificial intelligence algorithm for early sepsis diagnosis in the intensive care unit. *International Journal of Medical Informatics* 141 (2020), 104176. https://doi.org/10.1016/j.ijmedinf.2020.104176

[102] Hubert D. Zając, Dana Li, Xiang Dai, Jonathan F. Carlsen, Finn Kensing, and Tariq O. Andersen. 2023. Clinician-Facing AI in the Wild: Taking Stock of the Sociotechnical Challenges and Opportunities for HCI. *ACM Trans. Comput.-Hum. Interact.* 30, 2, Article 33 (mar 2023), 39 pages. https://doi.org/10.1145/3582430

[103] Massimo Zambon, Marcello Ceola, Roberto Almeida de Castro, Antonino Gullo, and Jean-Louis Vincent. 2008. Implementation of the Surviving Sepsis Campaign guidelines for severe sepsis and septic shock: We could go faster. *Journal of Critical Care* 23, 4 (2008), 455–460. https://doi.org/10.1016/j.jcrc.2007.08.003

[104] Dongdong Zhang, Changchang Yin, Katherine M. Hunold, Xiaoqian Jiang, Jeffrey M. Caterino, and Ping Zhang. 2021. An interpretable deep-learning model for early prediction of sepsis in the emergency department. *Patterns* 2, 2 (2021), 100196. https://doi.org/10.1016/j.patter.2020.100196

[105] 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. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. https://doi.org/10.1145/3351095.3372852

[106] Yuan Zhang, Chen Lin, Min Chi, Julie Ivy, Muge Capan, and Jeanne M. Huddleston. 2017. LSTM for septic shock: Adding unreliable labels to reliable predictions. In *2017 IEEE International Conference on Big Data (Big Data)*. 1233–1242. https://doi.org/10.1109/BigData.2017.8258049

[107] Zihao Zhu, Changchang Yin, Buyue Qian, Yu Cheng, Jishang Wei, and Fei Wang. 2016. Measuring Patient Similarities via a Deep Architecture with Medical Concept Embedding. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 749–758. https://doi.org/10.1109/ICDM.2016.0086

A FORMATIVE STUDY INTERVIEW SCRIPT

- **Question 1 - Background:** Can you tell us a bit more about your work, such as your job title, years of practice, main working context, etc.?
- **Question 2 - Experience of Sepsis Diagnosis:** Could you recall a recent sepsis encounter that stands out to you? When/where it happened? (Please do not mention any patient personally identifiable information)
- **Question 3 - Experience of Existing Epic Sepsis Module:** In the previous sepsis encounter, how did you use the sepsis diagnosis module in Epic EHR? Useful or not at all?
- **Question 4 - Attitudes towards AI Diagnosis:** How do you like the sepsis risks score predicted by AI model? [Do you want AI to assisted you?] [What information you wish AI could have provided you?]
- **Question 5 - Closing Question:** Is there anything else that you would like to share with us or any questions you have for us?

B LSTM MODEL PERFORMANCE

To demonstrate the performance of our chosen LSTM model, we have excerpted its performance at Figure 3 on two different cases from [104].

Case 1. In this case, patients are sampled from septic patients, and the goal is to see if a model can tell if a patient is likely to have high sepsis risk a few hours before the onset. For each patient, the patient records is split into 2 segments at the middle point, segment close to sepsis onset (= 4 hours) is labeled as 1, another segment (> 4 hours before sepsis onset) is labeled 0. We randomly pick either the former or latter segment to build the Case1 cohort. The introduction of case 1 is to measure the model in terms of time-sensitive prediction to ensure models are indeed clinically useful and relieve "warning fatigue" as alarm burden. Given patient records either from $T_{admission}$ to T_{middle} or from T_{middle} to $T_{onset} - 4$, our model is required to distinguish this 2 kinds of records.

Case 2. In this case, case and control segments are from different patients who have sepsis onset in the next 4 hours, as well as those who do not have sepsis. Given patient records from $T_{admission}$ to $T_{onset} - 4$, we are going to predict whether sepsis occurs in the following 4 hours.

C MODEL PARAMETERS

Sepsis onset risk prediction model. Following [104], we used PyTorch [61] to implement the LSTM-based sepsis risk prediction model, and the number of timesteps for LSTM was set to 100. The sizes of variable embedding vectors and hidden state vectors were set as 256. The number of layers in LSTM was set as 2. For evaluation, 80% of the data are used for training, and 10% for validation, 10% for testing. The results are displayed in Table 3.

Recommendation algorithm. Two thresholds th_s and th_e ($1 > th_s, th_e > 0$) are used to decide whether new laboratory values should be requested in subsubsection 4.3.2. The prediction threshold th_s was set as 0.5 to predict whether the patients will have sepsis onset in next 4 hours. The uncertainty threshold th_e was set as 0.25 to determine whether to collect more clinical variables (e.g., to request additional lab tests when Shannon entropy [71] uncertainty is bigger than 0.25). The results are displayed in Table 2.

Table 3: AUC accuracy scores of sepsis prediction tasks on DII-challenge dataset from [104].

Method	Case1	Case2	Average
MEWS	0.54	0.72	0.63
NEWS	0.52	0.72	0.62
SIRS	0.56	0.69	0.62
qSOFA	0.53	0.65	0.59
Logistic regression	0.89	0.79	0.84
Random forest	0.90	0.81	0.85
Gradient boosting trees	0.91	0.81	0.86
GRU	0.88	0.80	0.84
RETAIN [17]	0.90	0.80	0.85
Dipole [54]	0.90	0.81	0.86
LSTM [104]	**0.94**	**0.84**	**0.89**

D USER STUDY INTERVIEW SCRIPT

- **Question 1 - Attitude to AI Technology in Clinical Scenario:** Do you want any AI in your work?
- **Question 2 - Feedback on Our Initial Design (after showing the screenshot/demo of the UI):** How do you like the design?
- **Question 3 - Willingness to use:** Would you use such an AI system everyday?
- **Question 4 - Trust:** Would you trust it?
- **Question 5 - Improvement**: What improvement do you need?