Augmenting Clinical Decision-Making with an Interactive and Interpretable AI Copilot: A Real-World User Study with Clinicians in Nephrology and Obstetrics

Yinghao Zhu*
National Engineering Research Center for Software Engineering
Peking University
Beijing, China
School of Computing and Data Science
University of Hong Kong
Hong Kong, China
yhzhu99@gmail.com

Dehao Sui*
National Engineering Research Center for Software Engineering
Peking University
Beijing, China
dehaosui1@gmail.com

Zixiang Wang*
National Engineering Research Center for Software Engineering
Peking University
Beijing, China
wangzx@stu.pku.edu.cn

Xuning Hu
Department of Computer Science and Engineering
Hong Kong University of Science and Technology
Hong Kong, China
xuninghu@hkust-gz.edu.cn

Lei Gu
National Engineering Research Center for Software Engineering
Peking University
Beijing, China
leiguha99@gmail.com

Yifan Qi
National Engineering Research Center for Software Engineering
Peking University
Beijing, China
qiyifan2004@icloud.com

Tianchen Wu
Department of Obstetrics and Gynecology
Peking University Third Hospital
Beijing, China
wtc17@pku.edu.cn

Ling Wang
Department of Nephrology
Affiliated Xuzhou Municipal Hospital of Xuzhou Medical University
Jiangsu, China
lingw1228@sohu.com

Yuan Wei
Department of Obstetrics and Gynecology
Peking University Third Hospital
Beijing, China
weiyuanbysy@163.com

Wen Tang
Department of Nephrology
Peking University Third Hospital
Beijing, China
tanggwen@126.com

Zhihan Cui
Institute for Global Health and Development
Peking University
Beijing, China
zhihancui@pku.edu.cn

Yasha Wang
National Engineering Research Center for Software Engineering
Peking University
Beijing, China
wangyasha@pku.edu.cn

Lequan Yu
School of Computing and Data Science
University of Hong Kong
Hong Kong, China
lqyu@hku.hk

Ewen M Harrison
Centre for Medical Informatics
University of Edinburgh
Edinburgh, United Kingdom
ewen.harrison@ed.ac.uk

Junyi Gao
Centre for Medical Informatics
University of Edinburgh
Edinburgh, United Kingdom
Health Data Research UK
London, United Kingdom
junyii.gao@gmail.com

Liantao Ma†
National Engineering Research Center for Software Engineering
Peking University
Beijing, China
malt@pku.edu.cn

Abstract

Clinician skepticism toward opaque AI hinders adoption in high-stakes healthcare. We present AICare, an interactive and interpretable AI copilot for collaborative clinical decision-making. By analyzing longitudinal electronic health records, AICare grounds dynamic risk predictions in scrutable visualizations and LLM-driven

*Equal contribution.
†Corresponding author.

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

CHI '26, Barcelona, Spain

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791458

diagnostic recommendations. Through a within-subjects counterbalanced study with 16 clinicians across nephrology and obstetrics, we comprehensively evaluated AICare using objective measures (task completion time and error rate), subjective assessments (NASA-TLX, SUS, and confidence ratings), and semi-structured interviews. Our findings indicate AICare's reduced cognitive workload. Beyond performance metrics, qualitative analysis reveals that trust is actively constructed through verification, with interaction strategies diverging by expertise: junior clinicians used the system as cognitive scaffolding to structure their analysis, while experts engaged in adversarial verification to challenge the AI's logic. This work offers design implications for creating AI systems that function as transparent partners, accommodating diverse reasoning styles to augment rather than replace clinical judgment.

CCS Concepts

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

Keywords

Clinical Decision Support System; Human-AI Interaction; AI for Healthcare; Explainable AI

ACM Reference Format:

Yinghao Zhu, Dehao Sui, Zixiang Wang, Xuning Hu, Lei Gu, Yifan Qi, Tianchen Wu, Ling Wang, Yuan Wei, Wen Tang, Zhihan Cui, Yasha Wang, Lequan Yu, Ewen M Harrison, Junyi Gao, and Liantao Ma. 2026. Augmenting Clinical Decision-Making with an Interactive and Interpretable AI Copilot: A Real-World User Study with Clinicians in Nephrology and Obstetrics. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 26 pages. https://doi.org/10.1145/3772318.3791458

1 Introduction

The integration of Artificial Intelligence (AI) into the medical domain is currently defined by a sharp contrast. On one hand, algorithmic performance continues to break records in controlled experiments [19, 51, 65, 82, 84]. Recent Large Language Models (LLMs) and deep learning architectures often match or exceed human experts in diagnostic tasks ranging from radiology to medical licensing examinations [20, 50, 58]. On the other hand, the successful and sustained deployment of these systems into actual clinical practice remains rare [5, 47, 80]. This disconnect is widely termed the "last mile" problem of medical AI [48, 63]. It suggests that the primary barrier to adoption is no longer strictly computational, but related to clinical integration. Current failures in deployment often stem not from a lack of accuracy, but from a lack of clinical legitimacy, specifically the degree to which AI systems align with the complex and accountability-driven reasoning processes that define high-stakes medical practice [12, 64, 80].

Early paradigms in Clinical Decision Support Systems (CDSS) typically conceptualized AI as an authoritative predictor. These systems function as black boxes that deliver a final verdict, such as a risk score or a binary diagnosis, and expect passive acceptance from the human user [31, 81]. However, clinical decision-making is rarely a binary choice that can be automated. Instead, it is an iterative process involving hypothesis generation, evidence gathering, and cross-checking, where clinicians must weigh conflicting data points against patient history and clinical guidelines [60, 63, 81]. When AI systems bypass these intermediate cognitive steps to offer only a final conclusion, they may act as competitors to the clinician rather than supporters. This dynamic can trigger skepticism or resistance among experts who feel their professional agency is threatened [1, 15, 81]. Furthermore, the lack of transparency prevents clinicians from performing their ethical duty of verification, turning the clinician into a liable agent who is responsible for AI errors without possessing the means to interrogate or understand them [7, 31].

To bridge this gap between algorithms and practice, Human-Computer Interaction (HCI) research must pivot from designing automated predictors to designing AI copilots. These are collaborative partners designed not to dictate decisions, but to support the user's inquiry and process integration [63, 77, 78]. Unlike a static predictor, an AI copilot facilitates clinical legitimacy not by demanding blind trust, but by providing the interactive mechanisms for clinicians to triangulate evidence and actively interrogate the underlying data [13, 44, 62].

While the HCI community has begun to explore these themes, existing work remains limited by two critical factors. First, much of the current literature relies on formative interviews or the evaluation of non-interactive, static prototypes. There is a scarcity of studies that empirically evaluate how clinicians interact with fully functional systems capable of real-time data exploration in high-stakes environments. Second, prior work often treats clinicians as a monolithic group. This overlooks how the "last mile" challenges may diverge based on domain expertise. While theoretical frameworks warn of de-skilling for junior clinicians or algorithmic aversion from experts, there is limited empirical evidence demonstrating how interactive features might mitigate these risks by supporting distinct cognitive strategies for novices versus experts within the same system [7, 21, 72].

In this work, we present and evaluate **AICare**, an interactive and interpretable AI copilot. AICare translates the concept of clinical legitimacy by exposing its reasoning through dynamic risk trajectories, interactive feature-level attribution, and LLM-synthesized clinical narratives. Rather than delivering a static prediction, AICare grounds its dynamic risk predictions in interpretable visualizations, allowing the clinician to analyze the patient case collaboratively. AICare features the following core components designed to support clinical sensemaking:

(1) A **dynamic risk trajectory visualization** charting a patient's risk over time, providing a longitudinal narrative of their health journey.
(2) An **interactive list of critical risk factors** showing the clinical features most influential to a prediction, with on-demand drill-down into specific feature trends.
(3) An **LLM-driven diagnostic recommendation** that synthesizes the AI's key findings into a concise, clinical narrative.
(4) A **population-level indicator analysis** that contextualizes a patient's data by comparing their indicators against cohort-level trends.

To assess AICare's impact in a real-world clinical setting, we conducted a within-subjects user study with 16 clinicians utilizing the deployed system. We contend that one of the most complex and

critical challenges for clinical AI is the management of chronic conditions through long-term, longitudinal patient follow-up [16, 56], which involves reasoning over sparse, high-dimensional, and irregularly sampled time-series data [42]. Our preliminary stakeholder discussions further prioritized these high-risk contexts, citing the urgent need for support where error tolerance is minimal and the cognitive load of monitoring deterioration is highest. Therefore, we purposefully selected two representative specialties for our evaluation: nephrology, focusing on chronic disease management for patients with end-stage renal disease, and obstetrics, focusing on preterm birth risk assessment during prenatal care. Both scenarios involve high-stakes decisions where AI assistance could prevent severe adverse outcomes.

To guide our evaluation of AICare within these contexts and to unpack the resulting behavioral dynamics, we structure our investigation around the following three Research Questions (RQs):

- **RQ1:** How do clinicians perceive the utility and usability of AICare's interactive and interpretable modules when integrated into their diagnostic workflow?
- **RQ2:** What is the impact of AICare on clinicians' diagnostic efficiency, accuracy, and cognitive workload, and how does this impact vary across different clinical contexts and experience levels?
- **RQ3:** How do the interactive and interpretable features of AICare influence clinicians' trust in the AI's recommendations and their overall decision-making confidence?

Through task-based sessions and semi-structured interviews, our findings reveal that clinicians engaged with AICare not as an infallible authority, but as a cognitive partner for collaborative sensemaking. Quantitatively, we observed a significant reduction in perceived cognitive workload ($p = .023$) and a significant increase in diagnostic confidence ($p = .018$), while diagnostic accuracy remained comparable to the baseline. Regarding efficiency, although total task duration did not differ statistically, senior clinicians took slightly longer to complete tasks. This temporal trend aligns with granular interaction metrics that revealed a significant behavioral divergence ($p < .05$) based on expertise: senior clinicians engaged in more active data interrogation behaviors, acting out a process of adversarial verification. Conversely, junior clinicians exhibited lower interaction frequencies, leveraging the system primarily as a cognitive scaffold to structure their analysis.

Qualitatively, we found that trust was not passively granted but actively constructed through a process of human-AI alignment. Clinicians consistently used AICare's interactive features, such as tracing the patient's story along the risk trajectory and scrutinizing the feature importance list, to cross-reference the AI's rationale against their own mental model. This transformed the interaction from a simple consumption of a prediction into a collaborative dialogue. Crucially, moments of disagreement, where the AI's reasoning diverged from clinical intuition, did not necessarily erode trust. Instead, when supported by plausible evidence, they often became opportunities for critical reflection.

Furthermore, clinicians articulated a clear vision for the future of such tools. Beyond its current analytical capabilities, they expressed a strong desire for AI copilots that provide more context-aware and actionable recommendations, such as suggesting specific follow-up tests or linking counter-intuitive findings to relevant clinical guidelines. They envision a future where these systems are seamlessly integrated into the EHR, functioning as vigilant assistants that help them manage information overload to augment rather than replace their irreplaceable clinical judgment.

Overall, this paper makes the following contributions:

- The design and implementation of AICare, an interactive and interpretable AI copilot currently integrated into hospital information systems and deployed in clinical settings (accessible for demonstration at https://aicare.pages.dev/), translating a collaborative human-AI paradigm through scrutable, multi-level explanations.
- An empirical evaluation combining quantitative metrics and qualitative analysis with practicing clinicians in two distinct specialties and hospital tiers, providing insights into both shared principles and context-specific needs for AI adoption.
- A set of actionable design implications for creating AI copilots that foster a collaborative human-AI partnership, grounded in empirical evidence of distinct verification mechanisms: cognitive scaffolding for novices and adversarial verification for experts.

2 Related Work

Our research synthesizes three critical discourses in HCI and medical AI: the shift from static prediction to collaborative workflow integration, the reconceptualization of trust as an active verification process, and the mediating roles of clinical expertise and context in human-AI interaction.

2.1 Human-AI Collaboration in Clinical Workflows

The prevailing challenge in medical AI is no longer strictly regarding algorithmic capability, but rather the sociotechnical gap or "last mile" problem of deployment [5, 48, 80]. Historically, Clinical Decision Support Systems (CDSS) often functioned as rule-based alert systems, providing reminders or flagging potential drug interactions [6, 45, 61]. With the advent of machine learning, these systems evolved to offer predictive analytics, such as risk scores for patient deterioration or disease onset [55], with prominent examples like the Epic Sepsis Model providing a single risk score to alert clinicians [73].

However, recent scholarship argues that this model of AI as a diagnostic authority fundamentally misunderstands the nature of clinical work. Medical decision-making is inherently collaborative, iterative, and socially situated [63, 77, 81]. Yang et al. demonstrate that successful AI integration requires systems to be "unremarkable", implying they must be embedded seamlessly into the social fabric of decision-making rather than acting as disruptive external authorities that demand attention [77]. Similarly, ethnographic work has revealed that high-performing algorithms often fail because they lack the flexibility to handle the environmental realities of clinical workflows, such as poor lighting, internet latency, or patient hardship [5, 69]. These failures occur because rigid systems act as competitors to the clinician, challenging the authority of the clinician without understanding the full patient context [63, 81].

To address this misalignment, the field is moving toward human-AI teaming and copilot models. Zhang et al. propose systems that

support the intermediate stages of decision-making, such as hypothesis generation and data gathering, rather than attempting to automate the final verdict [81]. This aligns with findings that clinical usefulness is distinct from algorithmic accuracy; usefulness depends on the configurability of the system to local workflows, such as prioritizing worklists or drafting reports, rather than just diagnosing pathology [63, 78, 80]. Furthermore, Hussain et al. argue that unless interaction models shift from passive recommendation to active data exploration (e.g., exposing underlying data), clinicians become "liability sinks", responsible for AI errors they cannot understand or prevent because the system is inscrutable [31]. Our work answers these calls by evaluating a system designed specifically to support these intermediate reasoning steps, allowing clinicians to retain agency while leveraging AI for complex longitudinal data synthesis.

2.2 Interpretability and Trust Calibration in Medical AI

Traditionally, trust in medical AI has been modeled as a static attitude or a reliance score calibrated by the observed accuracy of the model [35]. However, recent literature frames trust as a dynamic, emergent state that is maintained through continuous interaction and verification [44, 66]. Yang et al. further attribute this difficulty to capability uncertainty, arguing that trust cannot be static because the AI functions as a living, sociotechnical system with evolving boundaries [76]. Systematic reviews posit that trust evolves through "contestability", the ability of the user to challenge the outputs of the system and probe its boundaries [66]. Similarly, recent frameworks describe trust as a mediator in a feedback loop (Input–Mediator–Output–Input), maintained through active monitoring processes rather than blind faith [44].

This shift reframes the role of Interpretable AI. The "black box" nature of many deep learning models is a significant barrier to their adoption in healthcare, a field where accountability and justification are paramount [59, 67]. Consequently, interpretability in AI has become a major focus [3, 29]. Common techniques include feature importance methods like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations), which highlight the factors driving a prediction [40, 54]. However, studies show that standard "feature importance" explanations often fail to improve decision outcomes because they increase cognitive load without improving the user's ability to detect errors [22, 46, 70]. Panigutti et al. address this via a progressive disclosure interface that presents natural language summaries before raw data to mitigate information overload [49]. Clinicians do not necessarily need mathematical transparency. They need clinical legitimacy, interpretability that aligns with medical ontology and clinical reasoning structures [12, 14, 64]. Researchers argue for legibility and cross-checking, where AI serves as one data signal among many that clinicians cross-reference against patient history [62, 63].

Moreover, trust construction is often an exercise in agency and alignment. Cai et al. found that pathologists need to understand the subjectivity and "medical personality" of the AI to treat it as a colleague rather than a tool [13]. Goh et al. further demonstrate that clinicians are willing to modify decisions based on AI recommendations if the system engages in a transparent reasoning process rather than static output [23]. Ideally, interactive systems should support forward reasoning, where the AI extends the user's thought process, engendering deeper engagement than those that simply provide a retrospective justification [50, 53]. Our study builds on this by identifying specific behavioral mechanisms (such as adversarial verification) by which experts construct trust in high-stakes environments.

2.3 Impact of Clinical Context and Expertise on AI Interaction

A critical but under-explored dimension of AI deployment is how clinical expertise modulates interaction and reliance. The one-size-fits-all approach to AI design is increasingly scrutinized as insufficient for diverse medical teams. Calisto et al. demonstrated that the optimal communication style of an AI varies by user: junior clinicians benefit from assertive guidance that provides structure, while experts prefer suggestive support that respects their autonomy and experience [14, 15].

For junior clinicians, AI introduces the risk of de-skilling or automation bias. Gaube et al. found that novices are highly susceptible to adhering to incorrect AI advice even when they claim to be skeptical, due to a lack of confidence in their own judgment [21]. Klingbeil et al. further illustrate that without domain expertise, users demonstrate "algorithm appreciation", favoring AI advice even when it contradicts contextual information [33]. Bienefeld et al. warn that without careful design, automation can atrophy the implicit knowledge required for holistic care [7, 8]. However, properly designed systems can act as cognitive scaffolding. In this mode, the AI helps novices structure their learning and organizes complex data without removing the cognitive effort required for mastery [63, 72].

Conversely, senior clinicians often approach AI with skepticism or algorithmic resistance. They frequently view it as a tool for efficiency rather than diagnosis [1]. They engage in adversarial behaviors, utilizing the AI to challenge their own hypotheses or to double-check against errors [23, 63]. While some studies suggest experts might reject AI that conflicts with their intuition, others suggest they are willing to modify decisions if the AI provides transparent, evidence-based counter-arguments that they can verify [23, 39]. Our work extends this literature by providing empirical evidence of these distinct behavioral modes within a single system evaluation. We demonstrate how AICare supports the learning of the novice via scaffolding while simultaneously withstanding the scrutiny of the expert.

3 The AICare System

AICare is a web-based, interactive, and interpretable AI copilot currently deployed and integrated into the daily clinical workflow of the participating hospitals to support clinicians in diagnostic risk assessment. Its architectural design was guided by a research-through-design approach, involving multiple iterative cycles of development and pilot deployment with senior nephrologists and obstetricians across six years.

During early stakeholder interviews, clinicians explicitly rejected models that output static risk scores, citing that such scores lacked the "narrative context" of disease progression. This specific clinical

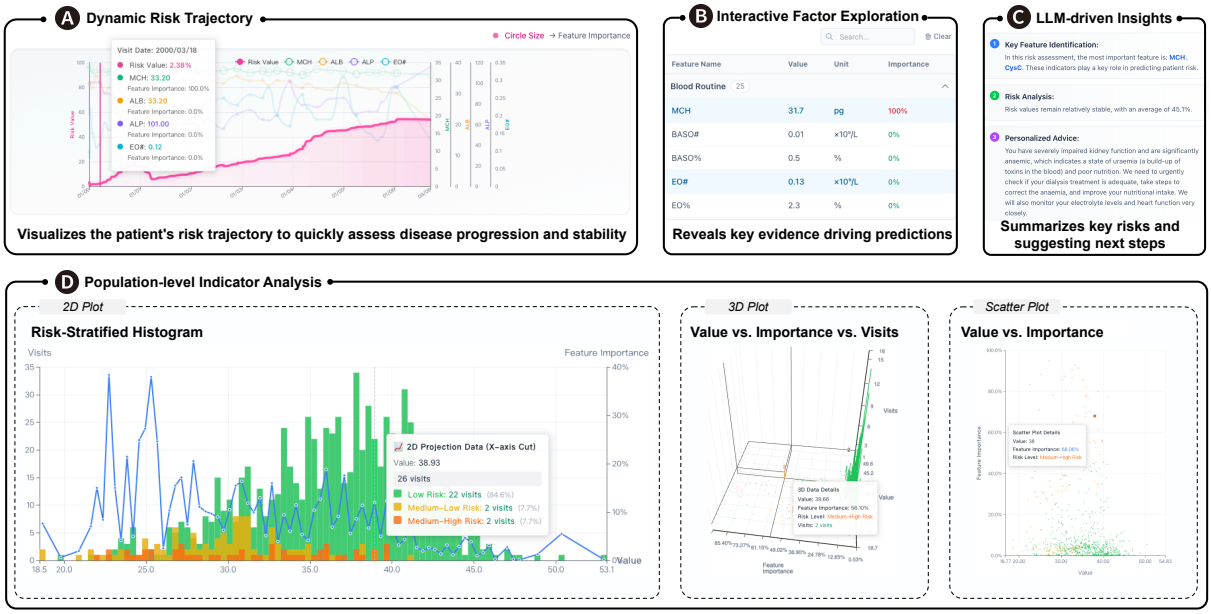

Figure 1: The AICare system interface, designed as an interactive and interpretable AI copilot. The dashboard features (A) a dynamic risk trajectory visualization over the patient's visits that externalizes memory-intensive trend analysis, (B) an interactive list of critical risk factors with their adaptive importance scores enabling immediate drill-down into evidence, (C) an LLM-driven diagnostic recommendation synthesizing key findings, and (D) a population-level indicator analysis providing cohort context for a selected feature. Interaction patterns reveal that these features support dual strategies: acting as cognitive scaffolding for novices to structure analysis, while enabling experts to perform "adversarial verification" of the model's logic.

requirement drove our underlying model selection. We moved away from standard MLPs (multilayer perceptrons) or tree-based models, which were initially considered, toward a time-aware architecture capable of visualizing longitudinal changes, a direct response to the clinicians' need to verify the "trajectory" rather than just the "state" of a patient.

3.1 Underlying Interpretable Predictive EHR Model

Feedback from our pilot deployments showed that clinicians do not rely on abstract feature weight; they require clinical legitimacy where they can intuitively verify that the model's logic aligns with the patient's history. To meet this requirement for longitudinal transparency, we adapted the ConCare architecture [41, 42] to our AICare model as the analytical engine. This design decision was made to address two specific characteristics of real-world EHR data identified by our clinical partners:

(1) **Handling irregular sampling via multi-channel extraction.** Clinical data is sparse and irregularly sampled. Unlike standard RNNs which assume fixed time steps, our implementation uses a bidirectional Gated Recurrent Unit (bi-GRU) to capture the varying time-lapses between visits. This ensures that the model treats a gap of one week differently than a gap of one year, while simultaneously grounding these dynamics in static baseline features (e.g., demographics) processed through a separate multilayer perceptron (MLP).

(2) **Dynamic feature attribution via attention mechanisms.** Clinicians emphasized that a biomarker (e.g., Creatinine) might be irrelevant at baseline but critical during an acute flare-up. To model this, we utilized an adaptive feature importance recalibration module. By treating the patient embedding as a "health context query", the model uses a multi-head self-attention mechanism to assign personalized importance scores that change over time. This allows the system to highlight different risk factors at different stages of the disease progression, matching the clinician's evolving focus.

This architecture allows AICare not only to predict a future risk (e.g., 1-year mortality) but also to generate a dynamic, personalized, and time-aware explanation for its prediction, which forms the basis for our interactive visualizations.

3.2 System Features

AICare's interface is organized into a primary individual patient dashboard (Figure 1A, B, C) and a secondary population analysis view (Figure 1D). Each component is designed to facilitate a step in the clinical reasoning process, from gaining a high-level overview to scrutinizing specific evidence and placing it in a broader context.

Dynamic risk trajectory visualization. Instead of presenting a single, static risk score, AICare visualizes the patient's predicted risk as a longitudinal trajectory over the course of their clinical visits (Figure 1A). This time-series graph provides clinicians with an immediate, holistic view of the patient's health status, revealing trends, periods of stability, or moments of rapid deterioration. Overlaid on

the main risk curve are the temporal plots of the most influential clinical indicators, such as hemoglobin or albumin. Clinicians can hover over any point on the trajectory to inspect a detailed tooltip showing the precise risk value and the corresponding values of key features for that specific visit. The size of each feature's marker on the graph visually encodes its importance, which is defined as its contribution to the final prediction, scaled as a value between 0 and 1. This visual encoding of feature weights makes the model's dynamic reasoning immediately apparent.

Interactive list of critical risk factors. This module (Figure 1B) offers a detailed, local explanation for the prediction corresponding to the latest clinical visit. It presents a comprehensive list of all clinical features, which are grouped by category (e.g., blood routine), sorted by their predictive importance, and fully searchable for ease of access. For each feature, the interface displays its measured value, unit, and calculated importance percentage. Selecting a feature in the interactive list instantly overlays its historical trend line onto the main risk trajectory graph (Figure 1A), allowing clinicians to visually correlate a specific indicator's fluctuations with changes in the patient's overall risk.

LLM-driven diagnostic recommendation. This component (Figure 1C) synthesizes the quantitative outputs of the model into a concise, narrative summary designed for clinical decision-making. The LLM-generated text is structured into three parts: "Key Feature Identification", which highlights the most critical indicators driving the risk assessment; "Risk Analysis", which offers a qualitative summary of the risk trajectory; and "Personalized Advice", which translates the findings into a preliminary clinical interpretation and suggests potential areas for further investigation. This feature bridges the gap between the model's data-driven findings and actionable clinical thought processes.

Population-level indicator analysis. To contextualize an individual's data, this view (Figure 1D) enables clinicians to compare a specific indicator against cohort-level trends derived by performing inference on a random subset of 100 samples and aggregating results across all patient visits. This process establishes a global baseline that visualizes the multi-dimensional relationships between three key variables: the feature's measured value, its calculated importance, and the predicted risk probability. These relationships are rendered in interactive formats, including 2D, 3D, and scatter plots, to reveal how specific feature magnitudes correlate with overall risk. By positioning the current patient within this distribution, clinicians can ascertain whether a flagged feature represents a consistent population-level risk pattern or an atypical anomaly, thereby strengthening the evidentiary basis of their decisions.

3.3 Implementation Details

Real-world clinical datasets and predictive tasks. To ensure the system's robustness and clinical validity, AICare was implemented and evaluated using three distinct real-world Electronic Health Record (EHR) datasets sourced from three different departments. These datasets cover two high-stakes specialties and are characterized by sparse, high-dimensional, and irregularly sampled time-series data (see Table 1 for statistics). Detailed inclusion/exclusion criteria and comprehensive feature statistics for all cohorts are provided in Appendix A and Appendix B.

- **Nephrology (ESRD Mortality Risk):** We utilized datasets from XY Hospital (a large regional general hospital) and BS Hospital (a top-tier national hospital). These collections track patients with End-Stage Renal Disease (ESRD) undergoing long-term dialysis. The data characteristics include static demographics (e.g., age, gender, primary disease) and longitudinal dynamic laboratory values (e.g., albumin, creatinine, potassium) recorded during varying follow-up intervals. The specific task is to predict the rolling risk of one-year all-cause mortality at each patient visit.
- **Obstetrics (Preterm Birth Risk):** We utilized a dataset from BC Hospital (a top-tier national hospital). This dataset comprises records of women with multiple gestation pregnancies. The data characterizes the prenatal journey through maternal demographics and time-varying clinical indicators (e.g., cervical length, uterine contractions, laboratory results). The predictive task is to assess the risk of spontaneous preterm birth (delivery < 37 weeks) at each prenatal check-up.

Table 1: Descriptive statistics of the three EHR datasets used to train the predictive models and source the patient cases for the study.

Statistic	XY-Nephrology	BS-Nephrology	BC-Obstetrics
Total Patients	1,404	341	5,315
Eligible Patients	1,224	280	3,577
Total Records Loaded	7,461	8,938	43,038
Avg. Visits (Timesteps)	6.46	19.20	12.03
Positive Outcome Ratio	26.80%	44.29%	54.12%
Feature Counts			
Static	7	23	29
Dynamic	33	66	50
Selected Cases for Study			
Total Selected	10	10	10
Positive Outcomes	5	3	5
Negative Outcomes	5	7	5

For each specialty (Nephrology and Obstetrics), we curated a set of 10 real, anonymized patient cases for the study. These cases were drawn from a held-out test set of their respective datasets to ensure the model had not been trained on them. We employed a stratified random sampling strategy based on ground-truth labels, thus including a mix of both positive and negative outcomes (e.g., mortality/survival, preterm/term birth). Furthermore, the selection included cases where the model's prediction was correct as well as cases where it was incorrect. This was done intentionally to observe how clinicians would interact with the system, calibrate their trust, and negotiate potential disagreements with the AI's recommendations. The ground truth outcome for each case was known to the researchers but not revealed to the participants during the study.

Validation of predictive performance benchmarks. To ensure that study participants interacted with a clinically competent system, we rigorously evaluated the underlying predictive models prior to deployment. Using a stratified 10-fold cross-validation strategy, the Nephrology model (predicting 1-year mortality) achieved an Area Under the Receiver Operating Characteristic (AUROC) of

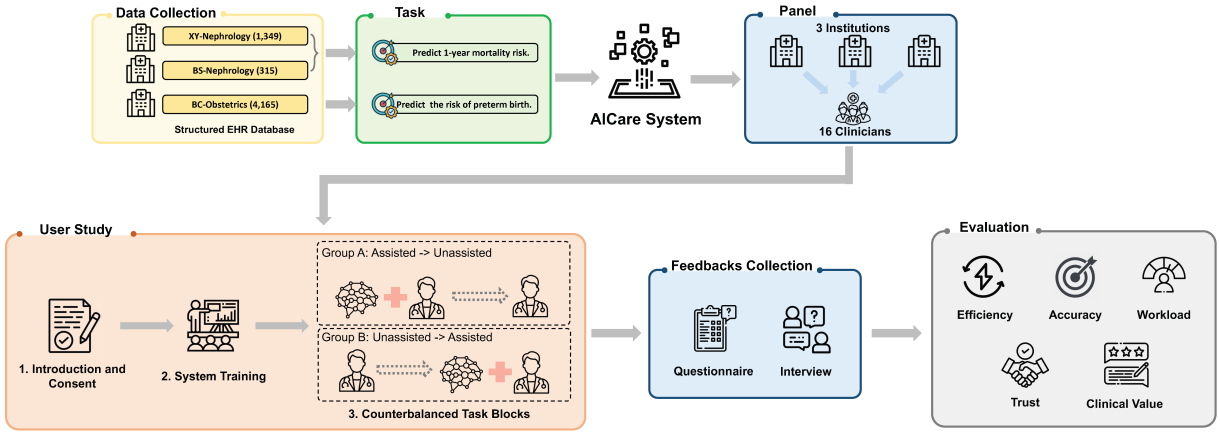

Figure 2: An overview of the user study methodology, from data collection to evaluation. We utilized structured EHR data from three institutions across two specialties (Nephrology and Obstetrics) to train the AICare system for two predictive tasks. The study involved 16 clinicians in a rigorous within-subjects, counterbalanced design. Participants completed risk assessment tasks under two conditions: AI-assisted (using AICare) and unassisted (a baseline). We collected quantitative and qualitative data through post-task questionnaires and semi-structured interviews to evaluate AICare's impact on efficiency, accuracy, cognitive workload, trust, and perceived clinical value. Quantitative results demonstrate that AICare significantly reduces cognitive workload by offloading the mental burden of synthesizing longitudinal data, without compromising diagnostic accuracy.

0.898 and an Area Under the Precision-Recall Curve (AUPRC) of 0.906, surpassing recent state-of-the-art benchmarks, such as the Transformer-based approach by Xu et al. (AUROC=0.877) [75]. Similarly, the Obstetrics model (predicting preterm birth) achieved an AUROC of 0.740 and AUPRC of 0.775, surpassing specialized baselines like Bacsaran et al. (AUROC=0.71) [4]. Establishing this high level of predictive fidelity was a prerequisite for the user study, ensuring that the risk trajectories and feature attributions presented to clinicians in AICare were derived from a robust and reliable analytical engine.

Training details for EHR models. Following the PyEHR toolkit [83], the modeling pipeline transforms these raw, longitudinal EHR datasets into calibrated, predictive models. The process includes feature engineering, handling of missing values using methods such as Last Observation Carried Forward (LOCF) [74] and median imputation, feature normalization, and balancing of class distributions through oversampling. Complete details regarding data preprocessing, model architecture, and hyperparameter configuration are provided in Appendix C.

The underlying EHR model for interpretability was implemented in PyTorch and PyTorch Lightning and was trained using the Adam optimizer, with an early stopping strategy based on the Area Under the Precision-Recall Curve (AUPRC) on a validation set to prevent overfitting. We employ a post-hoc optimization stage, which involves two key processes: first, we apply Temperature Scaling to calibrate the model's raw outputs, ensuring the predicted probabilities are more reliable and intuitive for clinical interpretation [24]. Second, we determine an optimal decision threshold by maximizing the F-beta score on the validation set, allowing us to tune the model's sensitivity and specificity to align with the specific diagnostic needs of each clinical scenario.

Development of AICare system. The AICare system is engineered as a decoupled web application. The backend is developed using FastAPI [52], a modern Python web framework chosen for its high performance and native support for asynchronous operations, which is crucial for handling model inference requests without blocking the user interface. Data persistence is managed by a SQLite database via the SQLAlchemy ORM. The backend loads pretrained AICare models and exposes a set of RESTful API endpoints that serve as the analytical core of the system. The frontend is a single-page application built with Vue 3 [79] and the ElementPlus UI library, ensuring a modular and responsive design. Interactive visualizations are rendered using Apache ECharts. For the LLM-driven diagnostic recommendation, the frontend directly queries the DeepSeek API using its official client, feeding it a structured prompt (detailed in Appendix D) containing the model's key findings to generate a concise narrative summary. To mitigate the risk of hallucinations, we employed a constraint-based prompting strategy that explicitly instructs the LLM to ground its narrative strictly in the provided quantitative evidence. Given that inconsistencies may still arise, the system relies on a human-in-the-loop validation model, allowing our study to empirically examine how clinicians perceive and scrutinize these AI-generated insights.

Hardware and software configuration. All EHR model training experiments were run on an Apple Mac Studio M3 Ultra with 512GB of RAM. The deployment of the AICare system was also based on the Mac Studio. The primary software stack comprises Python 3.12, PyTorch 2.6.0, PyTorch Lightning 2.5.1, and Transformers 4.50.0. Experiments were conducted between August 1, 2025, and September 1, 2025. The LLM-generated recommendations utilize the DeepSeek-V3 [37], DeepSeek-R1 [25], DeepSeek-V3.1 [18] model via its official API. Interview recordings were transcribed using

Gemini 2.5 Pro [17]. To strictly control for hallucinations or errors, the generated text was cross-validated by two researchers prior to inclusion in the qualitative analysis.

4 User Study Methodology

To investigate our research questions, we designed and conducted a within-subjects user study with a counterbalanced order of conditions (see Figure 2). This approach allowed us to rigorously evaluate the impact of AICare on clinicians' performance and perceptions while minimizing inter-participant variability and gathering rich qualitative insights into their experience. The study protocol received ethical approval from the institutions' research ethics committees.

4.1 Participants

We recruited a total of 20 medical professionals through professional contacts at three hospital sites representing two different tiers of the healthcare system. The participants were divided into two cohorts based on the study phase: a primary cohort for the main user study ($N = 16$) and a supplementary cohort for in-depth interviews ($N = 4$). All participants were compensated for their time. The study protocol received ethical approval from the Institutional Review Board (IRB) at each participating hospital (see Section 4.5).

Primary user study cohort ($N = 16$). Participants details for the main within-subjects experiment are summarized in Table 2. This group included:

- **XY Hospital (Nephrology):** Department of Nephrology, Affiliated Xuzhou Municipal Hospital of Xuzhou Medical University. A large regional hospital, from which we recruited 8 professionals, including physicians, nurses, and residents.
- **BS Hospital (Nephrology):** Department of Nephrology, Peking University Third Hospital. A top-tier national hospital, from which we recruited 5 professionals, including physicians, residents, and medical students.
- **BC Hospital (Obstetrics):** Department of Obstetrics and Gynecology, Peking University Third Hospital. A top-tier national hospital specializing in women and children's health, from which we recruited 3 attending or associate chief physicians.

Participants' clinical experience ranged from 1 to over 20 years (M around 11 with SD around 7). Their self-reported familiarity with AI-based CDSS varied (M around 2.8, SD around 1.0 on a 5-point scale), with senior clinicians often reporting less prior use than junior clinicians.

Supplementary interview cohort ($N = 4$). To complement the quantitative metrics with granular insights into clinical workflow integration, we recruited an additional group of 4 participants specifically for semi-structured interviews. These participants did not perform the timed diagnostic tasks but provided detailed narrative feedback on system usability and cognitive processing. Detailed demographics for this group are presented in Table 3.

4.2 Study Scenarios and Conditions

To evaluate generalizability across distinct medical contexts, we designed two study scenarios corresponding to the specific clinical tasks: (1) Nephrology, focusing on the 1-year mortality risk for patients with End-Stage Renal Disease (ESRD), and (2) Obstetrics, focusing on the risk of spontaneous preterm birth during prenatal care.

Participants analyzed 10 anonymized patient cases specific to their specialty. The study employed a within-subjects design with two experimental conditions to counterbalance the order of conditions and case sets to control for individual variability and order effects:

- **Unassisted (Control):** Participants reviewed a static, non-interactive digital summary of the patient's EHR data. This interface presented the same core information (demographics, visit history, lab results) available in AICare but without any predictive analytics or interactive visualizations.
- **AI-assisted (AICare):** Participants used the full AICare interface, which was populated with the same underlying patient data and our trained predictive models.

4.3 Procedure

The study session for each participant lasted approximately 60 to 75 minutes and was conducted one-on-one with a researcher in a quiet office. The procedure followed a structured protocol:

(1) **Introduction and Consent (approx. 5 min).** The researcher explained the study's purpose, obtained written informed consent, and the participant completed a background questionnaire (Appendix F.1).
(2) **System Training (approx. 10 min).** Participants received a standardized tutorial (following the script in Appendix E) on the AICare interface using a sample case not included in the main study tasks. This ensured a consistent level of proficiency with the system's features. During this session, we explicitly articulated the probabilistic nature of the AI and the possibility of errors, directing participants' attention to the interface's disclaimer that all AI outputs are provided as "preliminary clinical interpretations" requiring human verification.
(3) **Counterbalanced Task Blocks (approx. 40 min).** Participants were randomly assigned to one of the following two groups (AI-first or Human-first):
    - *Group A (n=9, AI-first):* Analyzed 5 cases using AICare, then 5 cases in the unassisted condition.
    - *Group B (n=7, Human-first):* Completed the tasks in the reverse order.

    For each case, the participant's task was to review the patient's record and provide a risk assessment (e.g., "What is the likelihood of mortality in the next year?") on a 4-point Likert scale (Low, Low-Medium, Medium-High, High), along with their confidence in that assessment on a 5-point Likert scale (see Appendix F.2). We logged the time taken for each case.
(4) **Workload Assessment.** Immediately after completing each block of 5 cases, participants filled out the NASA-Task Load Index (NASA-TLX) questionnaire to assess their perceived cognitive workload for that condition [26].
(5) **Post-Task Questionnaires (approx. 10 min).** After completing all 10 cases, participants filled out the System Usability Scale (SUS) [36] and the Trust in Automation scale for the AICare system. In addition, participants completed a "AICare Feature

Table 2: Participant demographics and background ($N = 16$). Participants were from three hospital sites (XY, BC, BS). Condition: "AI-first" participants used the AI system before their own analysis; "Human-first" participants did the reverse. "AI Fam." (AI Familiarity): Self-reported on a 5-point Likert scale (1: Completely Unfamiliar, 5: Expert). Exp. (Experience).

ID	Role	Exp. (Yrs)	Age Group	Gender	Department	Condition	Interviewed	AI Fam.	Prior AI Use
XY-01A	Assoc. Chief Physician	16-20	40-49	F	XY-Nephrology	AI-first	No	2	Occasionally
XY-02B	Nurse	11-15	30-39	F	XY-Nephrology	Human-first	Yes	2	Occasionally
XY-03A	Attending Physician	6-10	30-39	F	XY-Nephrology	AI-first	Yes	3	Never
XY-04B	Assoc. Chief Physician	16-20	40-49	M	XY-Nephrology	Human-first	Yes	2	Never
XY-05A	Chief Physician	16-20	40-49	M	XY-Nephrology	AI-first	No	2	Never
XY-06B	Assoc. Chief Nurse	11-15	30-39	F	XY-Nephrology	Human-first	Yes	2	Occasionally
XY-07A	Resident Physician	1-5	30-39	M	XY-Nephrology	AI-first	Yes	3	Frequently
XY-08B	Chief Physician	>20	>50	F	XY-Nephrology	Human-first	Yes	3	Never
BS-01A	Medical Student	1-5	<30	F	BS-Nephrology	AI-first	No	3	Frequently
BS-02B	Medical Student	1-5	<30	F	BS-Nephrology	AI-first	No	3	Frequently
BS-03A	Resident Physician	1-5	<30	F	BS-Nephrology	Human-first	No	3	Daily
BS-04B	Chief Physician	>20	40-49	F	BS-Nephrology	Human-first	Yes	5	Never
BS-05A	Intern	1-5	<30	F	BS-Nephrology	AI-first	No	5	Occasionally
BC-01A	Attending Physician	6-10	30-39	F	BC-Obstetrics	AI-first	Yes	3	Frequently
BC-02B	Assoc. Chief Physician	11-15	30-39	F	BC-Obstetrics	Human-first	No	4	Occasionally
BC-03A	Attending Physician	6-10	30-39	F	BC-Obstetrics	AI-first	Yes	3	Frequently

Table 3: Demographics of the additional participant group (N=4) interviewed for the discussion analysis. "Condition" indicates the counterbalanced order: "AI-first" participants used the AICare system prior to the conventional tabular interface, while "Table-first" participants followed the reverse order. "AI Fam.": Self-reported AI familiarity on a 1–5 scale.

ID	Role	Exp. (Yrs)	Age Group	Gender	Department	Condition	Interviewed	AI Fam.	Prior AI Use
BC-A1	Resident Clinician	1-5	<30	F	BC-Obstetrics	Table-first	Yes	3	Never
BC-A2	Attending Clinician	1-5	30-39	M	BC-Obstetrics	AI-first	Yes	5	Daily
BC-A3	Attending Clinician	6-10	30-39	M	BC-Obstetrics	Table-first	Yes	4	Frequently
BC-A4	Medical Student	<1	<30	F	BC-Obstetrics	AI-first	Yes	4	Never

Feedback" form (Appendix F.6), where they rated the specific usefulness of AICare's individual modules.

(6) **Semi-Structured Interview (approx. 10-15 min).** The session concluded with an audio-recorded interview with participants who consented (n=9). The interview, detailed in Appendix F.7, explored their overall impressions of AICare, focusing on trust, usability, workflow integration, and the system's clinical value.

4.4 Measures and Data Analysis

Our study employs a combination of quantitative and qualitative analyses to provide a comprehensive evaluation. We quantitatively analyze performance metrics and survey responses, and qualitatively analyze data from semi-structured interviews.

Quantitative measures. We collected data across five key dimensions: efficiency, accuracy, workload, usability, and trust. Each measure is detailed below.

- **Efficiency.** We measure efficiency as the task completion time in seconds for each case assessment. The timer starts when a participant is first presented with a case and stops upon completing a valid diagnosis and submitting their confidence rating. Time spent in discussions with the research team for procedural clarifications or to provide feedback on the user interface is excluded from this measurement. This ensures that the data reflects only the time dedicated to the assessment task itself.
- **Accuracy.** Participants assessed patient risk using a 4-point Likert scale (A: 0–25%, B: 25–50%, C: 50–75%, D: 75–100%). To compute accuracy, we first binarize these assessments into a low-risk category (A and B) and a high-risk category (C and D). This binarized assessment is then compared against the patient's actual binary outcome (ground truth) to determine correctness. We calculate the accuracy for each participant as the percentage of correct assessments under each condition (unassisted and AI-assisted).
- **Workload.** We assess cognitive workload using the NASA Task Load Index (NASA-TLX) [26, 27] (see Appendix F.3). Participants rated their perceived workload across six dimensions (mental demand, physical demand, temporal demand, performance, effort, and frustration) on a continuous scale (e.g., "How much mental and perceptual activity was required?" for mental demand, and "How insecure, discouraged... did you feel?" for frustration). In our analysis, we use the raw TLX score, which is the unweighted average of the six ratings, resulting in a composite workload score from 0 to 100 for each condition.
- **Usability.** We evaluate the usability of the AICare system with the System Usability Scale (SUS) [10] (Appendix F.4), a standardized 10-item questionnaire with a 5-point Likert scale ranging

from 1 ("Strongly disagree") to 5 ("Strongly agree"). Items include statements such as "I found the system unnecessarily complex" and "I thought the system was easy to use". The final SUS score is calculated by converting responses according to standard procedure: for positively worded items (1, 3, 5, 7, 9), we subtract 1 from the score; for negatively worded items (2, 4, 6, 8, 10), we subtract the score from 5. The sum of these converted scores is then multiplied by 2.5, yielding a final usability score from 0 to 100 [2].

- **Trust.** We measure participants' trust in the AI system using the Trust in Automation Scale [32] (Appendix F.5). This scale consists of 12 items rated on a 7-point Likert scale (1="Strongly disagree", 7="Strongly agree"), including questions such as "The system is deceptive" and "I am confident in the system." Several items (1, 4, 6, 9, 11) are negatively worded and are reverse-scored by subtracting their value from 8. The final trust score is the average of all 12 item scores after reversal, resulting in a value between 1 and 7.
- **Interaction Behaviors.** To verify the depth of cognitive engagement, we analyzed screen recordings to extract granular interaction metrics. We quantified specific actions as proxies for verification intensity: (1) *List paging:* Frequency of scrolling or paging through the risk factor list, serving as a proxy for information seeking beyond the top-ranked features; (2) *Curve hovering:* Frequency of hovering over specific time points on the risk trajectory. In time-series visual analytics, hovering is often correlated with detailed value inspection and local trend verification [28]; (3) *Cross-view comparison:* Frequency of moving the cursor between the historical trend view and current status view, indicating an active process of synthesizing longitudinal context with cross-sectional evidence.

Qualitative measures. The nine semi-structured interviews (guiding questions provided in Appendix F.7) were transcribed and analyzed using a reflexive thematic analysis approach [9]. To ensure accuracy, the AI-generated transcripts were manually cross-validated by two researchers against the original audio recordings prior to analysis. The two researchers independently coded the transcripts, identified initial themes, and then collaboratively met to discuss, merge, and refine these themes into a final thematic structure that captured the core aspects of the clinicians' experience with AICare.

Statistical analysis. We analyzed the collected data using statistical methods appropriate for the experimental design. For the within-subjects factors (efficiency, accuracy, and workload), we conducted a repeated measures analysis of variance (RM-ANOVA). In cases where the data violated the assumption of normality, we employed the Aligned Rank Transform (ART) for non-parametric factorial analysis [71], a robust method for examining main effects and interactions. The assumption of sphericity was assessed using Mauchly's test, and when violated, Greenhouse–Geisser (GG) corrections were applied to adjust the degrees of freedom; all reported p-values for within-subject effects reflect the GG-adjusted tests when applicable. For the between-subjects factor of task order (AI-assisted first vs. unassisted first), we used an independent samples t-test to examine its effect on measures such as trust. For all analyses, we report effect sizes (partial eta-squared, η_p^2, for ANOVA; Cohen's d for t-tests). Post-hoc pairwise comparisons were adjusted using Holm's correction for multiple comparisons. All statistical analyses were performed in R, interfaced through the `rpy2` library in Python, utilizing the `ARTool` [71] and `afex` packages.

4.5 Ethical Considerations

Our research was conducted in full compliance with the ACM Policy on Research Involving Human Participants and Subjects and the ethical principles of the Declaration of Helsinki. All study procedures, including participant recruitment and informed consent, were reviewed and approved by the Medical Science Research Ethics Committees of the participating hospitals prior to study commencement.

We obtained written informed consent from all clinician participants. The consent process ensured they understood the study's purpose, the voluntary nature of their participation, and their right to withdraw at any time. All collected data, including audio recordings and survey responses, were anonymized to protect participant confidentiality. Participants were compensated for their professional time.

5 Results Analysis

We present our findings organized by our research questions, integrating quantitative and qualitative results to provide a comprehensive picture. The qualitative insights derived from our semi-structured interviews were synthesized using reflexive thematic analysis to capture the nuances of clinician interaction. Figure 3 presents the resulting thematic map, organizing findings into three high-level categories aligned with our research questions: (1) the perception of utility and usability, highlighting the balance between high functional value and potential visual clutter; (2) the impact on efficiency, accuracy and workflow, emphasizing the system's role in stimulating clinical reasoning versus confirming known risks; and (3) the influence on trust, distinguishing between drivers of confidence (such as interactive verification) and sources of skepticism (such as logical contradictions or blind spots).

5.1 Analysis to RQ1

AICare is perceived as useful but visually demanding, with variations across hospital tiers. Overall, participants rated AICare's features as highly useful for clinical practice, with a mean usefulness score of 79.69 (SD = 14.70, Table 4). The dynamic risk trajectory visualization (M = 82.81) and interactive list of critical risk factors (M = 81.25) were deemed the most valuable components. Clinicians appreciated the system's ability to narrate a patient's health history. For instance, a nurse from the regional hospital (XY-02B) noted, "Seeing the risk trend over time is far more valuable than a single score." Similarly, a resident from the national hospital (BS-03A) found the trend charts "easy to understand", noting that observing a drop in hemoglobin helped quickly associate it with clinical events like bleeding. Beyond individual diagnosis, participants envisioned specific integration points for AICare within the broader workflow: XY-02B described it as "a small alarm bell" to enhance patient communication during follow-ups, while XY-08B viewed it as an auxiliary tool for clinical rounds.

However, System Usability Scale (SUS) score revealed a noticeable divide (Table 5). While the overall score was 63.91 (marginally

Figure 3: Thematic map of clinician interview findings. The hierarchical structure organizes key themes and sub-themes derived from semi-structured interviews. Findings are grouped by each of the three core research questions (RQs) and are supported by representative quotes from participants. Qualitative analysis confirms that transparency amplifies trust when the reasoning is clinically plausible, but accelerates rejection when the AI contradicts established medical logic.

Table 4: Descriptive statistics of perceived usefulness scores for system features. F1: Dynamic Risk Trajectory Visualization, F2: Interactive List of Critical Risk Factors, F3: LLM-driven Diagnostic Recommendation, F4: Population-level Indicator Analysis. Scores were rated on a scale from 0 (not at all useful) to 100 (extremely useful).

Category	Feature	N	Mean±SD	Median	Min	Max
Overall Score	Usefulness Score	16	79.69±14.70	81.25	50.00	100.00
Individual Feature	F1	16	82.81±11.97	75.00	75.00	100.00
	F2	16	81.25±19.36	75.00	50.00	100.00
	F3	16	78.12±20.16	75.00	25.00	100.00
	F4	16	76.56±24.95	75.00	25.00	100.00
By Department	XY-Nephrology	8	85.16±12.91	81.25	68.75	100.00
	BS-Nephrology	5	78.75±14.39	81.25	62.50	100.00
	BC-Obstetrics	3	66.67±15.73	68.75	50.00	81.25
By Seniority	Senior clinicians	9	75.69±13.42	75.00	50.00	100.00
	Junior clinicians	7	84.82±15.67	81.25	62.50	100.00

acceptable), clinicians at the regional hospital (XY) rated it higher ($M = 72.19$) than those at national centers (BS: $M = 56.00$; BC: $M = 55.00$). This performance gap implies that the system's utility may be stratified by clinical expertise, institutional workflow, case complexity, or the specific model's utility. In qualitative feedback, while clinicians found the system relatively easy to learn (Q7 $M = 80.00$), many found the interface unnecessarily complex (Q2 $M = 45.31$, reversed) and cumbersome (Q8 $M = 35.94$, reversed). For instance, several participants (XY-07A, BS-03A, BS-04B, BC-01A) criticized the dynamic risk plot when multiple indicators were overlaid, describing the chart as "messy", "too narrow" or "overwhelming", leading one senior nurse (XY-02B) to suggest that the core curve should be amplified to avoid information overload. Notably, while interaction patterns were consistent across domains, the visualization required domain-specific adaptation, such as anchoring the timeline to gestational weeks in obstetrics versus calendar dates in nephrology.

The LLM-driven insights provide efficiency but lack domain specificity. The LLM-driven diagnostic recommendation (Figure 1C) received mixed feedback ($M = 78.12$). It was praised by some as a "quick consult note" (BC-01A) that streamlined the initial review. However, domain-specific limitations were evident. In obstetrics, where protocols are strictly defined by gestational age, BC-03A found the text recommendations "not well-integrated with clinical practice", preferring explicit risk percentages at specific timepoints (e.g., 28 or 34 weeks) over generic text. Furthermore, BS-04B described the suggestions as "boilerplate" and "not very professional", indicating that for the LLM to be truly effective in specialized centers, it must move beyond general summaries to offer guideline-specific, actionable interventions.

Table 5: Descriptive statistics of the System Usability Scale (SUS) scores. Both the total score and individual item scores are presented on a 0-100 scale. Item labels Q1-Q10 correspond to the individual questions of the SUS, with (R) indicating a reverse-scored item.

Category	Item	N	Mean±SD	Median	Min	Max
Overall Score	SUS Score	16	63.91±13.72	67.50	45.00	90.00
Individual Item	Q1: Frequent Use	16	71.88±17.97	75.00	25.00	100.00
	Q2: Unnecessarily Complex (R)	16	45.31±20.85	50.00	25.00	100.00
	Q3: Easy to Use	16	73.44±21.35	75.00	25.00	100.00
	Q4: Need Technical Support (R)	16	51.56±29.54	37.50	25.00	100.00
	Q5: Functions Well Integrated	16	71.88±15.48	75.00	50.00	100.00
	Q6: Too Much Inconsistency (R)	16	45.31±27.72	50.00	0.00	100.00
	Q7: Quick to Learn	15	80.00±19.36	75.00	50.00	100.00
	Q8: Awkward/Cumbersome (R)	16	35.94±25.77	25.00	0.00	100.00
	Q9: Felt Confident	16	60.94±20.35	50.00	25.00	100.00
	Q10: Needed to Learn a Lot (R)	16	35.94±27.34	25.00	0.00	100.00
By Department	XY-Nephrology	8	72.19±12.57	68.75	50.00	90.00
	BS-Nephrology	5	56.00±10.69	50.00	45.00	67.50
	BC-Obstetrics	3	55.00±9.01	52.50	47.50	65.00
By Seniority	Senior clinicians	9	67.22±15.43	67.50	47.50	90.00
	Junior clinicians	7	59.64±10.75	67.50	45.00	70.00

Contextual modules serve research needs over daily clinical routine. The population-level indicator analysis (Figure 1D) was identified as having a distinct "research value" (BC-03A), scoring moderately high in perceived usefulness ($M = 76.56$). Residents and clinicians (XY-07A, XY-03A) praised its ability to contextualize individual values, establish local safety thresholds (e.g., defining green/red zones), and generate hypotheses for scientific inquiry. However, the practical application of this module in a fast-paced setting was questioned; obstetrician BC-01A noted it had a "high learning curve" and would likely be used for "model validation by engineers" or retrospective research rather than point-of-care decisions. This delineates a clear separation of concerns: while the risk trajectory directly supports the active clinical workflow, the population analysis serves to support the broader scientific workflow.

RQ1's Key Findings and Implications:

(1) Clinicians prioritize AI's explanatory power over its predictive output. The high utility ratings for interpretable features (risk trajectory, feature importance), especially among junior clinicians, underscore a demand for tools that support, rather than bypass, clinical reasoning.

(2) Usability is highly context-dependent and challenged by information density. The system's moderate overall SUS score, suggests that a one-size-fits-all design is insufficient. A primary design challenge is managing information density to avoid cognitive friction.

(3) Generative recommendations require strict grounding. The LLM module was valued for efficiency but criticized for lack of domain specificity and occasional hallucination. To be viable, generative features in high-stakes settings must move beyond general summarization to guideline-anchored and strictly verifiable output.

5.2 Analysis to RQ2

AICare did not statistically alter overall efficiency, but qualitative patterns suggest expertise-based usage differences. Our study reveals a nuanced impact of AICare on diagnostic efficiency. To capture precise interaction behaviors and timing, we relied on high-resolution screen recordings. Valid screen recordings were available for a subset of 12 participants (9 senior clinicians and 3 junior clinicians), determined by participant consent and data completeness. Consequently, the following analyses on efficiency and interaction metrics are restricted to this subgroup.

Overall, the average time taken to complete a case was comparable between the AI-assisted and unassisted conditions (Table 6, $p = .774$). When stratifying participants by clinical experience, we observed a potential trend: junior clinicians (≤ 5 years of experience) showed a decrease in average task time of approximately 32%. Although this result did not reach statistical significance ($p = .130$) given the subgroup sample size ($N = 3$), the substantial reduction suggests a distinct behavioral adaptation compared to senior clinicians (> 5 years of experience), who took slightly longer when using the system.

Table 6: Comparison of task completion time (in seconds) between the AI-assisted and unassisted conditions. Data analysis is based on $N = 12$ participants (9 Seniors, 3 Juniors) for whom complete, valid screen recordings were available. Statistical significance was assessed using repeated-measures ANOVA (RM-ANOVA). Note that for the junior clinicians group, a large effect size (partial $\eta_p^2 = 0.664$) was observed, indicating a substantial difference in task duration despite the statistical constraints of the subgroup sample size ($N = 3$).

	AI-assisted		Unassisted		Statistical Comparison		
Group	N	Mean±SD	N	Mean±SD	F	p	η_p^2
Overall	12	117.68±63.24	12	113.25±54.63	0.087	.774	0.012
By Department							
XY-Nephrology	8	134.57±67.71	8	124.85±57.80	0.601	.463	0.041
BC-Obstetrics	3	95.87±40.83	3	96.07±53.74	0.250	.667	<.001
By Seniority							
Senior clinicians	9	131.96±61.23	9	114.09±60.39	1.449	.263	0.209
Junior clinicians	3	74.87±57.75	3	110.73±42.58	6.250	.130	0.664

To understand the behavioral mechanisms driving this temporal difference, we analyzed granular interaction logs (Table 7). The data uncovers a statistically significant disparity in verification strategies: senior clinicians engaged in substantially more rigorous data interrogation than juniors. Specifically, seniors performed significantly more curve hovering ($F = 41.33, p < .001, \eta_p^2 = 0.805$) and cross-view comparison ($F = 7.07, p = .024, \eta_p^2 = 0.414$). This indicates that experts did not passively accept the AI's risk prediction; rather, they actively manipulated the timeline to inspect specific data points and frequently moved their focus between historical trends and current evidence. In contrast, junior clinicians showed significantly lower interaction frequencies (e.g., mean list paging 0.93 vs. 2.04 for seniors). This behavioral divergence suggests distinct reliance strategies: while experts engaged in "adversarial verification" by physically interrogating the data to test the model's logic, junior clinicians appeared to utilize the system's

pre-computed synthesis as a "cognitive scaffold," relying on the provided visual structure to frame their analysis rather than performing granular, bottom-up validation.

Table 7: Comparison of interaction behaviors between senior and junior clinicians. Data analysis is based on $N = 12$ participants (9 Seniors, 3 Juniors) for whom complete, valid screen recordings were available. An asterisk (*) indicates a statistically significant result. ($p < .05$)

	Senior ($N = 9$)	Junior ($N = 3$)	Statistical Comparison		
Metric	**Mean±(SD)**	**Mean±(SD)**	***F***	***p***	η_p^2
List Paging	2.04±(0.56)	0.93±(0.09)	9.615	**.011***	0.490
Curve Hovering	2.13±(0.28)	0.60±(0.43)	41.328	**< .001***	0.805
Cross-View Comparison	1.69±(0.69)	0.40±(0.57)	7.067	**.024***	0.414

Regarding accuracy, we evaluated diagnostic performance on a subset of cases ($N = 8$ participants from XY-Nephrology) where robust ground truth outcomes were available and the sample size allowed for valid comparison. As shown in Table 8, while accuracy improved marginally in the AI-assisted condition (+10.0%), the difference was not statistically significant ($p = .619$).

Table 8: Comparison of diagnostic accuracy between AI-assisted and unassisted conditions across different clinician groups. Analysis is restricted to the XY-Nephrology cohort ($N = 8$) where the sample size and ground truth distribution supported a valid comparative analysis. The AI model's benchmarked performance on the test set (10 cases from XY-Nephrology) was: Accuracy 80.0%, Precision 66.7%, Recall 66.7%, Specificity 85.7%, AUROC 0.643, and AUPRC 0.542. Repeated-measures ANOVA (RM-ANOVA) was used to compare accuracy between conditions. An improvement with a "+" indicates higher accuracy in the AI-assisted condition.

Group	N	AI-assisted (%)	Unassisted (%)	↑ (%)	p
Overall clinicians	8	72.5±21.2	62.5±29.2	+10.0	.619
By Seniority					
Senior clinicians	5	80.0±20.0	56.0±32.9	+24.0	.253
Junior clinicians	3	60.0±20.0	73.3±23.1	-13.3	.580

Qualitative feedback corroborates that these interaction patterns stem from distinct cognitive strategies. Junior clinicians tended to use AICare as a cognitive scaffold that structured and accelerated their analytical process. For example, resident XY-07A noted, "The overall risk curve is very important... I will first look at it to judge the patient's overall risk for the next year." Obstetricians BC-01A, when dealing with the specific, time-bound risk of preterm birth, reported that the tool was particularly efficient for "screening scenarios in outpatient clinics", where rapid risk stratification is paramount. For them, the system provided a clear starting point and a structured overview. Senior clinicians, conversely, engaged in a more deliberate dialogue with the AI, using the extra time for verification and criticism. For instance, BS-04B, a chief physician, spent considerable time analyzing the "cluttered" charts to scrutinize the system's logic against her own. This suggests that for experts, the AI introduces a "second opinion" verification step that, while valuable for safety, is inherently additive to the temporal workflow.

The system substantially reduced perceived cognitive workload across all clinician groups. A key benefit of AICare was its ability to alleviate the cognitive burden associated with analyzing complex patient cases. Participants reported a significant reduction in overall cognitive workload in the AI-assisted condition compared to the unassisted condition, as measured by the NASA-TLX scale (Table 9, $p = .023$). This reduction was particularly pronounced among junior clinicians, who reported a statistically significant decrease in workload ($p = .028$).

Table 9: Comparison of cognitive workload, measured by the NASA-TLX average score (scale 0-100), between the AI-assisted and unassisted conditions. Lower scores indicate a lower perceived workload. Statistical significance was assessed using repeated-measures ANOVA (RM-ANOVA). An asterisk (*) indicates a statistically significant result. ($p < .05$)

	AI-assisted		Unassisted		Statistical Comparison		
Group	**N**	**Mean±SD**	**N**	**Mean±SD**	**F**	**p**	η_p^2
Overall	16	41.55±15.54	16	47.49±12.51	6.385	.023*	0.182
By Department							
XY-Nephrology	8	31.85±8.45	8	40.33±11.01	1.833	.218	0.274
BS-Nephrology	5	52.13±16.92	5	55.70±11.43	4.500	.101	0.220
BC-Obstetrics	3	49.78±15.21	3	52.89±8.40	0.333	.622	0.041
By Seniority							
Senior clinicians	9	38.35±14.12	9	44.39±12.73	1.195	.306	0.134
Junior clinicians	7	45.67±17.40	7	51.48±11.92	8.292	.028*	0.370

The system achieves this by offloading the demanding task of synthesizing and interpreting large volumes of longitudinal EHR data. The integrated visualizations, such as the dynamic risk trajectory (Figure 1A) and the interactive feature importance list (Figure 1B), provide a pre-digested analytical summary. This allows clinicians to focus their cognitive resources on higher-level clinical reasoning, interpretation, and verifying the AI's logic, rather than on the manual and time-consuming process of data aggregation. Participant BS-03A described the visualization of risk trends as "very intuitive", contrasting it with the high mental effort of reading tabular data. Even among senior clinicians who were critical of the interface's visual density, the reduction in mental demand was acknowledged.

RQ2's Key Findings and Implications:

(1) Efficiency impact is experience-dependent with maintained accuracy, driven by distinct interaction behaviors. While overall time remained stable, interaction logs reveal that senior clinicians engage in significantly more active verification than juniors. This suggests AICare acts as a cognitive scaffold for novices but as a peer for experts.

(2) Cognitive load is significantly reduced. By synthesizing complex longitudinal data into an intuitive dashboard, AICare offloads the cognitive burden of data aggregation. This frees up clinicians' mental resources to focus on higher-level patient care and complex decision-making.

5.3 Analysis to RQ3

Trust is actively constructed through verification, driving decision-making confidence. Overall, clinicians rated AICare with a moderately high degree of trust ($M = 5.18/7$, $SD = 0.78$; see Table 10). However, our qualitative findings indicate that clinician trust is not a static attribute passively granted to the AI, but an active process of verification. Quantitative results show that while diagnostic accuracy did not significantly improve, clinicians' decision-making confidence increased significantly ($p = .018$, Table 11).

Table 10: Descriptive statistics for the trust in AI score, measured on a 7-point Likert scale, across different participant groups. An independent samples t-test between the "AI First" and "Human First" groups found no significant difference in trust scores. ($t(14) = 0.016, p = .987, d = 0.008$)

Category	Group	N	Mean±SD
Overall	All Participants	16	5.18±0.78
By Department			
	XY-Nephrology	8	5.54±0.57
	BS-Nephrology	5	5.07±0.77
	BC-Obstetrics	3	4.42±0.88
By Seniority			
	Senior clinicians	9	5.37±0.95
	Junior clinicians	7	4.94±0.42
By AI Usage Order			
	AI First	9	5.19±0.77
	Human First	7	5.18±0.84

Qualitative feedback reveals that this boost stems from the system's scrutability, which allows users to validate their own intuitions against the AI's evidence. Chief physician XY-08B described this dynamic as "a process of verification and comparison", where she used the tool not to generate an answer, but to confirm if her line of thought was consistent with the data. This verification was often facilitated by specific interactive mechanisms; as attending clinician XY-03A explained, "When I click on a point in time, the system immediately shows all the key indicator values... This is very useful, it makes everything clear at a glance, and that enhances trust." This suggests that transparency was functionally defined by participants as verifiability. Trust could be reinforced when interactive features allowed clinicians to confirm that the AI's logic was grounded in clinically sound evidence, specifically the alignment of feature importance with established pathophysiological mechanisms. Similarly, resident BS-03A noted that she approached the system with a "natural sense of trust", but it was the visual confirmation of the risk trajectory that solidified her assurance. Consequently, even when the AI's prediction differed slightly from the clinician's initial view, this transparency reduced subjective uncertainty. As obstetrician BC-03A noted, the "system's explanation function helps in understanding the difference", enabling her to make a final decision with conviction. Thus, AICare helps transform the black box of prediction into a verifiable process, allowing clinicians to decouple their confidence from the model's raw accuracy.

Table 11: Comparison of clinician diagnostic confidence between the AI-assisted and unassisted conditions. Confidence was measured on a 5-point Likert scale (1 = Not confident at all, 5 = Very confident), where higher scores indicate greater confidence. Repeated-measures ANOVA (RM-ANOVA) was used to compare scores between the two conditions. An asterisk (*) indicates a statistically significant result. ($p < .05$)

	AI-assisted		Unassisted		Statistical Comparison		
Group	**N**	**Mean±SD**	**N**	**Mean±SD**	**F**	**p**	η_p^2
Overall	16	3.71±0.77	16	3.29±0.87	7.018	.018*	0.253
By Department							
XY-Nephrology	8	3.90±0.67	8	3.63±0.65	5.453	.052	0.336
BS-Nephrology	5	3.72±0.80	5	3.00±1.17	2.087	.222	0.284
BC-Obstetrics	3	3.20±1.06	3	2.87±0.76	4.000	.184	0.641
By Seniority							
Senior clinicians	9	3.73±0.89	9	3.51±0.81	6.400	.035*	0.347
Junior clinicians	7	3.69±0.65	7	3.00±0.92	4.703	.073	0.324

Transparency is a double-edged sword: Plausibility determines trust. The interactive nature of AICare magnified the impact of the AI's reasoning quality. When the system's logic was transparent but divergent from clinical expectation, it often sparked productive negotiation. For example, one attending clinician (XY-03A) noted that while she typically focused on indicators like albumin and potassium, the AI's emphasis on sodium and PTH "opened up my thinking." She explained that such a discrepancy would prompt her to investigate the literature, thereby using the AI as a tool to challenge her own "fixed and limited" diagnostic habits. However, this same transparency made trust highly fragile when the revealed logic was clinically implausible. Because clinicians could see the "why", errors were immediately glaring. Chief physician BS-04B expressed explicit distrust when the system prioritized Cystatin C in a context contradicting conventional knowledge. Associate chief nurse XY-06B also expressed skepticism regarding the model's feature weighting, explicitly stating she doubted whether "the model might be overemphasizing" the importance of uric acid across different patient contexts. More critically, XY-08B highlighted that logical contradictions, such as the LLM flagging "digestive system disease" when the feature value was zero, were fatal to trust: "When the conclusion contradicts clinical common sense, the doctor cannot trust

it." Furthermore, concerns about data integrity, such as BC-01A's skepticism regarding imputation methods for missing values, surfaced precisely because the system made these data dependencies visible. This demonstrates that while interpretability can bridge the gap between human and AI, it requires the AI's reasoning to be not just explainable, but clinically robust; otherwise, visibility accelerates the loss of trust. Furthermore, the boundaries of trust were reinforced by what resident XY-07A termed AI's "observational blind spots", where the system cannot capture non-quantifiable bedside data, such as subtle changes in mental state or edema, requiring the clinician to bridge the gap between algorithmic probability and patient reality.

RQ3's Key Findings and Implications:

(1) Verification drives confidence. Clinicians do not blindly trust the AI; they test it. The significant rise in diagnostic confidence stems from the ability to interactively verify the AI's evidence (e.g., via drill-down interactions), which validates their own clinical intuition.

(2) Interpretability as a double-edged sword. Transparency amplifies both trust and distrust. While it allows clinicians to catch errors (safety), it also means that "hallucinations" or illogical feature weights (e.g., zero-value risks) cause immediate, catastrophic drops in trust.

(3) Plausibility enables learning. Disagreement between Human and AI fosters learning only when the AI's rationale is clinically plausible. Without grounding in standard medical logic or external evidence, novel insights are dismissed as algorithmic artifacts.

(4) Data integrity and observational blind spots delimit trust. Transparency regarding data dependencies exposes vulnerabilities that can erode trust. Furthermore, clinicians explicitly withhold trust for non-quantifiable physical signs and subjective patient states that remain invisible to the model.

6 Discussion

We interpret our findings by synthesizing the quantitative results from the primary study with the narrative insights derived from the supplementary interview cohort ($N = 4$, described in Table 3). Specifically, we discuss the shift from tabular cognition to visual recognition, the divergence in verification behaviors between experts and novices, and the resulting implications for designing practical clinical AI. As synthesized in Figure 4, our investigation follows a progression from inquiry to implication.

6.1 Does AICare Really Outperform Tabular Systems? The Role of Interactive Verification

AICare outperforms the tabular status quo not merely by providing predictions, but by altering the cognitive architecture of the risk assessment task. Quantitatively, this is evidenced by a significant reduction in perceived cognitive workload compared to baseline tabular systems, with improvements ranging from 11.3% to 36.1% across participants (BC-A1: 32.3%, BC-A2: 36.1%, BC-A3: 11.3%, BC-A4: 12.5%). Theoretical frameworks in cognitive engineering and medical informatics suggest that standard hospital information systems force clinicians to perform high-effort "mental simulation" on matrices of static values to reconstruct patient history [43, 81]. By externalizing this computation into the dynamic risk trajectory, AICare replaces the memory-intensive task of integrating row-and-column data with a high-bandwidth visual channel. This aligns with findings by Zając et al., who argue that clinically useful AI must move beyond accuracy to support the temporal and contextual realities of workflow integration [80]. Consequently, clinicians can perceive critical trends (e.g., deterioration slopes) as pre-attentive attributes rather than calculated derivatives.

Qualitatively, the superiority of the AI copilot stems from its ability to mitigate the information overload inherent in high-dimensional tabular datasets [34]. The cognitive burden of raw matrices was explicitly highlighted, as standard systems make it "difficult to process every detail" (BC-A2). By directly rendering the "trend of indicators" (BC-A1), the system replaces memory-intensive search with pre-attentive recognition. This shift transforms verification from a disjointed task into an immediate investigation; whereas tabular interfaces force users to hunt for correlations, the copilot makes the specific "contribution" of each feature accessible on demand (BC-A3). Furthermore, to mitigate the risk of missing signals in dense data, the system enables users to "verify AI warnings" by quickly locating indicator changes in key gestational weeks, acting as a safety net for subtle patterns (BC-A4) [69].

6.2 Why and When Clinicians Challenge AI Reasoning?

Clinicians, particularly experts, challenge AI reasoning because trust is constructed through a process of adversarial verification rather than passive acceptance. Our interaction metrics revealed that senior clinicians engaged in high-frequency data interrogation (e.g., curve hovering) not due to inefficiency, but because they treat AI outputs as hypotheses requiring rigorous evidence. This behavior mirrors the "negotiation" phase described by Sivaraman et al., where clinicians actively probe AI outputs to establish reliability before acceptance [59]. Qualitative feedback elucidates that this scrutiny arises from a need to map AI logic onto internal mental models; trust is effectively established only when the feature importance "conforms to medical logic" (BC-A3), a phenomenon consistent with the "contestability" requirements for trust in clinical CDSS [66]. Furthermore, static risk probabilities fail to satisfy expert requirements for causal explanation. Clinicians cannot judge urgency from a score alone without visualizing the "change and slope" of indicators (BC-A2). Therefore, the challenge behavior is a mechanism to expose the underlying evidence, transforming the AI from a black-box oracle into a scrutable peer that must demonstrate its reasoning process to be trusted. This aligns with frameworks on meaningful human oversight, where interpretable features act as instruments for verification rather than blind reliance [23, 44]. By enabling active auditing, the system preserves professional accountability, ensuring the clinician remains the responsible agent.

Active interrogation of AI outputs is triggered specifically during complex decision-making contexts involving "grey area situations" or safety-critical redundancy checks. While junior clinicians utilize

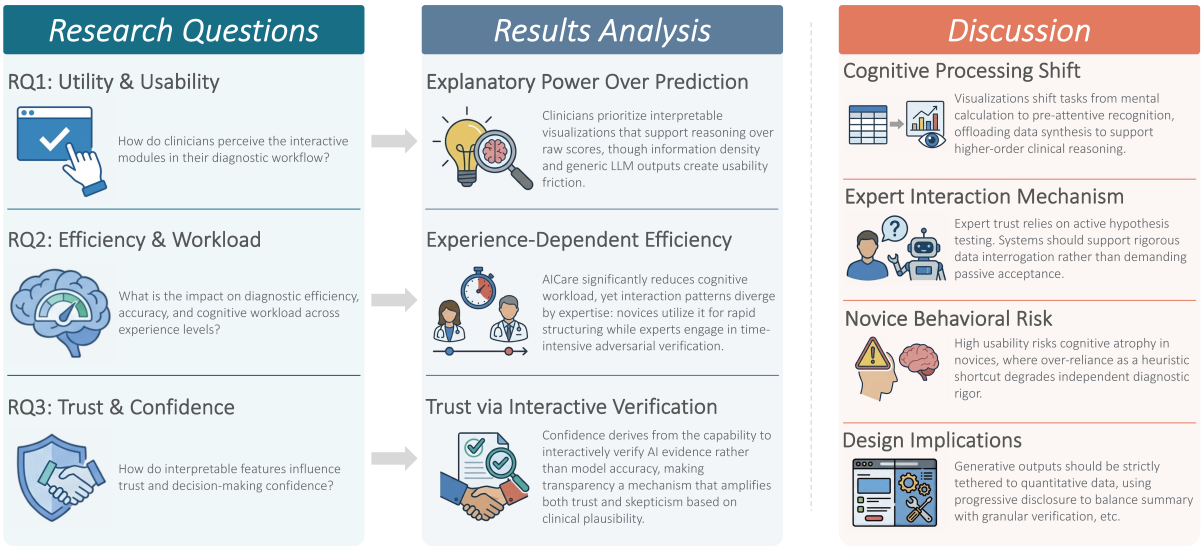

Figure 4: Study framework mapping the progression from research questions to results and discussion. The diagram illustrates three core pathways: the shift from prediction to explanatory power (RQ1), the divergence of interaction strategies based on clinical expertise (RQ2), and the active construction of trust through verification (RQ3), culminating in implications for cognitive processing and system design.

the system as a cognitive scaffold to combat the "mental inertia" of processing massive datasets (BC-A2), senior clinicians activate their critical verification primarily when facing ambiguity. In optional treatment scenarios where the path forward is unclear, the system serves as a crucial "second medical opinion," a role that Lu et al. found significantly improves decision quality when the AI provides distinct, evidence-based reasoning [39]. Conversely, in routine checks, the motivation shifts from decision support to safety assurance; the core value becomes "checking for omissions and filling gaps" rather than replacing human judgment (BC-A1). Thus, clinicians challenge the AI most intensely not when the case is obvious, but when the decision boundaries are blurred or when the cost of a missed indicator is unacceptably high [72].

6.3 Risk of Algorithmic Over-reliance and Passive Acceptance in Junior Clinicians

Our observations point to a potential vulnerability in the deployment of clinical AI systems: the susceptibility of novice clinicians to automation bias. While the system successfully mitigates the "mental inertia" caused by massive data overload, this cognitive relief creates a dangerous pathway toward passivity. For junior clinicians, who are still developing their internal diagnostic models, the high usability of AICare may inadvertently encourage them to treat the system's output as a verified conclusion rather than a hypothesis to be tested [21].

The danger is that the AI functions not as a pedagogical scaffold, but as a heuristic shortcut, leading to premature closure where the clinician bypasses the essential step of independent verification. This dynamic threatens to fundamentally erode the diagnostic rigor required for high-stakes medicine. By viewing the AI primarily as a "magical tool" for efficiency rather than a peer for consultation, junior clinicians risk a gradual de-skilling of their critical faculties. If the visualization of risk trajectories is accepted without scrutiny, the clinician is reduced from an active analyst to a passive supervisor, creating a scenario where they are liable for algorithmic hallucinations they are no longer equipped to catch [33]. Consequently, the very features designed to augment clinical capacity may paradoxically atrophy the critical thinking necessary for safe, independent practice unless deliberate cognitive forcing functions are introduced [11].

6.4 Design Implications for Next Generation Clinical AI Copilots

To bridge the gap between algorithmic output and clinical cognition, future systems could address distinct behavioral modes through the following strategies:

(1) **Strengthen the coupling between generative summaries and quantitative evidence.** To mitigate the risk of hallucinations and support the verification behaviors observed in our study, interfaces should tether LLM-generated narratives directly to their quantitative sources. For instance, future systems could make the provenance of each claim explicit by visually linking narrative points to the underlying data, instead of presenting the summary in isolation. This approach would enable clinicians to seamlessly validate the AI's qualitative interpretation against the ground-truth records.

(2) **Manage cognitive load through context-aware progressive disclosure.** To combat information overload without hiding critical data, systems should adopt a layered information architecture. Instead of presenting all feature contributions simultaneously, the interface should initially present only the

"synthesis layer" (overall risk trend and top 3 drivers). Detailed feature lists and population analytics should be revealed only on demand or triggered by specific anomaly detections. For junior clinicians, this acts as cognitive scaffolding; for experts, it reduces visual noise, allowing them to drill down only when the synthesis contradicts their intuition.

(3) **Prioritize clinical sense-making to drive everyday acceptance.** While interactive verification builds initial trust, sustained acceptance in daily workflows relies on shifting the focus from reasoning about the AI to reasoning about the patient. Our findings suggest that if an AI system demands constant, high-friction auditing, it risks becoming a burden rather than a support. To ensure long-term adoption, design strategies must go beyond merely exposing feature weights. Instead, they must align with the clinician's natural sense-making process, which involves synthesizing fragmented data into a coherent patient narrative. By positioning the AI as a partner that actively supports the construction of the clinical picture rather than a logic puzzle to be solved, the system transitions from a liable tool to an accepted professional asset.

(4) **Treat algorithmic competence as a design prerequisite.** Interactive transparency acts as a magnifying glass: it builds trust when the model is robust but accelerates rejection when logic is flawed. Although techniques like retrieval-augmented generation (RAG) can mitigate hallucinations by grounding generative outputs in established clinical guidelines, they cannot compensate for fundamental algorithmic deficiencies. If the underlying retrieval algorithm fails to surface specific, high-confidence evidence, the design must suppress generative output rather than risking plausible but false fabrications. Ultimately, high-fidelity predictive performance remains the non-negotiable foundation upon which safe interaction is built.

6.5 Limitations, Future Work and Ethical Implications

Our study has limitations characteristic of early-stage sociotechnical evaluations, specifically a modest sample size and reliance on retrospective data which constrain statistical generalizability. Technologically, the simultaneous visualization of high-dimensional longitudinal data introduced visual complexity, occasionally increasing cognitive load, while the probabilistic nature of LLMs poses safety risks regarding plausible yet unverified hallucinations. These findings highlight the need for optimized visual hierarchies and rigorous grounding mechanisms to balance information density with clinical safety.

Future work will leverage the ongoing deployment to conduct longitudinal analysis of usage patterns over extended periods. We aim to implement interactive counterfactual analysis where clinicians can manually modulate input variables, such as lab values, to immediately observe shifts in the risk trajectory, allowing doctors to stress-test clinical hypotheses directly. Additionally, we will develop adaptive interfaces that tailor guidance to user expertise, providing pedagogical scaffolding for novices and rapid verification for experts.

Broadly, AICare advances a paradigm shift from automation to cognitive augmentation. However, deploying such systems necessitates strict safeguards against automation bias and liability. Passive acceptance by novices risks atrophying independent diagnostic skills, while generative hallucinations threaten to burden clinicians with liability for unforeseen errors. Ethical design must therefore enforce cognitive friction to compel active verification, ensuring the system functions as a tool for scrutiny rather than a shortcut to premature closure. Ultimately, this work provides a blueprint for responsible AI integration where algorithmic intelligence empowers rather than displaces human judgment in high-stakes environments.

7 Conclusion

Our work reframes the challenge of medical AI adoption from one of algorithmic accuracy to sociotechnical collaboration. By evaluating AICare across distinct clinical specialties, we demonstrate that interactive interpretability empowers clinicians to actively verify rather than passively accept algorithmic risk assessments. The significant reduction in cognitive workload and the parallel increase in diagnostic confidence confirm that well-designed AI copilots can manage the information overload of longitudinal care without compromising human agency. Crucially, the distinct interaction patterns observed between experts and novices reveal that transparent systems could serve dual roles: functioning as platforms for adversarial verification by senior clinicians to validate model behavior, and as pedagogical scaffolds for junior clinicians to structure their reasoning. We conclude that the future of responsible healthcare lies in integrated systems that prioritize scrutability over simple automation, fostering a symbiotic partnership where computational power amplifies rather than replaces the irreplaceable nuance of clinical expertise.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62402017, 82470774, 82561128248, 82504426), Research Grants Council of Hong Kong (27206123, 17200125, C5055-24G, T45-401/22-N), Hong Kong Innovation and Technology Fund (GHP/318/22GD), Beijing Natural Science Foundation (QY25224, L244063, L244025), Peking University (Clinical Medicine Plus X Pilot Project 2024YXXLHGG007; "TengYun" Clinical Research Program TY2025015). Liantao Ma was supported by Beijing Traditional Chinese Medicine Science and Technology Development Fund (BJZYZD-2025-13), Beijing Municipal Health Commission Research Ward Excellence Clinical Research Program (BRWEP2024W032150205), and Xuzhou Scientific Technological Projects (KC23143). Junyi Gao acknowledges the receipt of studentship awards from the Health Data Research UK-The Alan Turing Institute Wellcome PhD Programme in Health Data Science (grant 218529/Z/19/Z) and Baidu Scholarship. We also thank the Apple Education Team for their support in hardware environment deployment.

References

[1] Anne Kathrine Petersen Bach, Trine Munch Nørgaard, Jens Christian Brok, and Niels Van Berkel. 2023. "If I had all the time in the world": Ophthalmologists' perceptions of anchoring bias mitigation in clinical AI support. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–14.

[2] Aaron Bangor, Philip T Kortum, and James T Miller. 2008. An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction* 24, 6 (2008), 574–594.

[3] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. 2020. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. *Information Fusion* 58 (2020), 82–115.

[4] Ezgi Başaran, Atakan Tanaçan, Nihat Farisoğullari, Zahid Ağaoğlu, Osman Onur Özkavak, Özgür Kara, and Dilek Şahin. 2025. The role of the lower uterine segment thickness in predicting preterm birth in twin pregnancies presenting with threatened preterm labor. *Journal of Perinatal Medicine* 53, 1 (2025), 39–47.

[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 (CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–12.

[6] Eta S. Berner (Ed.). 2016. *Clinical Decision Support Systems*. Springer International Publishing, Cham.

[7] Nadine Bienefeld, Emanuela Keller, and Gudela Grote. 2024. Human-AI teaming in critical care: A comparative analysis of data scientists' and clinicians' perspectives on AI augmentation and automation. *Journal of Medical Internet Research* 26 (2024), e50130.

[8] Nadine Bienefeld, Emanuela Keller, and Gudela Grote. 2025. AI interventions to alleviate healthcare shortages and enhance work conditions in critical care: qualitative analysis. *Journal of Medical Internet Research* 27 (2025), e50852.

[9] Virginia Braun and Victoria Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3, 2 (Jan. 2006), 77–101.

[10] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[11] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 5, CSCW1 (2021), 1–21.

[12] 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*. 1–19.

[13] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. "Hello AI": uncovering the onboarding needs of medical practitioners for human-AI collaborative decision-making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[14] Francisco Maria Calisto, João Maria Abrantes, Carlos Santiago, Nuno J Nunes, and Jacinto C Nascimento. 2025. Personalized explanations for clinician-AI interaction in breast imaging diagnosis by adapting communication to expertise levels. *International Journal of Human-Computer Studies* 197 (2025), 103444.

[15] Francisco Maria Calisto, João Fernandes, Margarida Morais, Carlos Santiago, João Maria Abrantes, Nuno Nunes, and Jacinto C Nascimento. 2023. Assertiveness-based agent communication for a personalized medicine on medical imaging diagnosis. In *Proceedings of the 2023 CHI conference on human factors in computing systems*. 1–20.

[16] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks.

[17] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261* (2025).

[18] DeepSeek. 2025. DeepSeek-V3.1 Release. https://api-docs.deepseek.com/news/news250821

[19] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. 2019. A Guide to Deep Learning in Healthcare. *Nature Medicine* 25, 1 (2019), 24–29.

[20] Junyi Gao, Yinghao Zhu, Wenqing Wang, Zixiang Wang, Guiying Dong, Wen Tang, Hao Wang, Yasha Wang, Ewen M Harrison, and Liantao Ma. 2024. A comprehensive benchmark for COVID-19 predictive modeling using electronic health records in intensive care. *Patterns* 5, 4 (2024).

[21] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 4, 1 (2021), 31.

[22] 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.

[23] Ethan Goh, Bryan Bunning, Elaine C Khoong, Robert J Gallo, Arnold Milstein, Damon Centola, and Jonathan H Chen. 2025. Physician clinical decision modification and bias assessment in a randomized controlled trial of AI assistance. *Communications Medicine* 5, 1 (2025), 59.

[24] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*. PMLR, 1321–1330.

[25] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* (2025).

[26] Sandra G Hart. 2006. NASA-task load index (NASA-TLX); 20 years later. In *Proceedings of the human factors and ergonomics society annual meeting*, Vol. 50. Sage publications Sage CA: Los Angeles, CA, 904–908.

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

[28] Jeffrey Heer and Ben Shneiderman. 2012. Interactive dynamics for visual analysis: A taxonomy of tools that support the fluent and flexible use of visualizations. *Queue* 10, 2 (2012), 30–55.

[29] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and Explainability of Artificial Intelligence in Medicine. *WIREs Data Mining and Knowledge Discovery* 9, 4 (2019), e1312.

[30] Yingxiang Huang, Wentao Li, Fima Macheret, Rodney A Gabriel, and Lucila Ohno-Machado. 2020. A tutorial on calibration measurements and calibration models for clinical prediction models. *Journal of the American Medical Informatics Association* 27, 4 (2020), 621–633.

[31] Muhammad Hussain, Ioanna Iacovides, Tom Lawton, Vishal Sharma, Zoe Porter, Alice Cunningham, Ibrahim Habli, Shireen Hickey, Yan Jia, Phillip Morgan, et al. 2024. Development and translation of human-AI interaction models into working prototypes for clinical decision-making. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference*. 1607–1619.

[32] Jiun-Yin Jian, Ann M Bisantz, and Colin G Drury. 2000. Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics* 4, 1 (2000), 53–71.

[33] Artur Klingbeil, Cassandra Grützner, and Philipp Schreck. 2024. Trust and reliance on AI—An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior* 160 (2024), 108352.

[34] Andre W Kushniruk and Vimla L Patel. 2004. Cognitive and usability engineering methods for the evaluation of clinical information systems. *Journal of biomedical informatics* 37, 1 (2004), 56–76.

[35] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[36] James R. Lewis. 2018. The system usability scale: Past, present, and future. *International Journal of Human–Computer Interaction* 34, 7 (2018), 577–590.

[37] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437* (2024).

[38] Josep Lluis Berral, Oriol Aranda, Juan Luis Dominguez, and Jordi Torres. 2021. Distributing Deep Learning Hyperparameter Tuning for 3D Medical Image Segmentation. *arXiv e-prints* (2021), arXiv–2110.

[39] Zhuoran Lu, Dakuo Wang, and Ming Yin. 2024. Does more advice help? the effects of second opinions in AI-assisted decision making. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–31.

[40] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* 30 (2017).

[41] Liantao Ma, Chaohe Zhang, Junyi Gao, Xianfeng Jiao, Zhihao Yu, Yinghao Zhu, Tianlong Wang, Xinyu Ma, Yasha Wang, Wen Tang, Xinju Zhao, Wenjie Ruan, and Tao Wang. 2023. Mortality Prediction with Adaptive Feature Importance Recalibration for Peritoneal Dialysis Patients. *Patterns* 4, 12 (2023), 100892.

[42] Liantao Ma, Chaohe Zhang, Yasha Wang, Wenjie Ruan, Jiangtao Wang, Wen Tang, Xinyu Ma, Xin Gao, and Junyi Gao. 2020. Concare: Personalized clinical feature embedding via capturing the healthcare context. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 833–840.

[43] John D Manning, Beatriz E Marciano, and James J Cimino. 2013. Visualizing the data–using lifelines2 to gain insights from data drawn from a clinical data repository. *AMIA summits on translational science proceedings* 2013 (2013), 168.

[44] Melanie J McGrath, Andreas Duenser, Justine Lacey, and Cecile Paris. 2025. Collaborative human-AI trust (CHAI-T): A process framework for active management of trust in human-AI collaboration. *Computers in Human Behavior: Artificial Humans* (2025), 100200.

[45] Randolph A. Miller. 1994. Medical Diagnostic Decision Support Systems—Past, Present, and Future: A Threaded Bibliography and Brief Commentary. *Journal of the American Medical Informatics Association* 1, 1 (1994), 8–27.

[46] Myura Nagendran, Paul Festor, Matthieu Komorowski, Anthony C Gordon, and Aldo A Faisal. 2023. Quantifying the impact of AI recommendations with explanations on prescription decision making. *NPJ Digital Medicine* 6, 1 (2023), 206.

[47] National Academy of Medicine; The Learning Health System Series. 2023. *Artificial Intelligence in Health Care: The Hope, the Hype, the Promise, the Peril.* National Academies Press (US), Washington (DC).

[48] Tariq Osman Andersen, Francisco Nunes, Lauren Wilcox, Elizabeth Kaziunas, Stina Matthiesen, and Farah Magrabi. 2021. Realizing AI in healthcare: challenges appearing in the wild. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*. 1–5.

[49] Cecilia Panigutti, Andrea Beretta, Daniele Fadda, Fosca Giannotti, Dino Pedreschi, Alan Perotti, and Salvatore Rinzivillo. 2023. Co-design of human-centered, explainable AI for clinical decision support. *ACM Transactions on Interactive Intelligent Systems* 13, 4 (2023), 1–35.

[50] Niroop Channa Rajashekar, Yeo Eun Shin, Yuan Pu, Sunny Chung, Kisung You, Mauro Giuffre, Colleen E Chan, Theo Saarinen, Allen Hsiao, Jasjeet Sekhon, et al. 2024. Human-algorithmic interaction using a large language model-augmented artificial intelligence clinical decision support system. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–20.

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

[52] Sebastián Ramírez. 2018. FastAPI.

[53] Leon Reicherts, Zelun Tony Zhang, Elisabeth von Oswald, Yuanting Liu, Yvonne Rogers, and Mariam Hassib. 2025. AI, help me think—but for myself: Assisting people in complex decision-making by providing different kinds of cognitive support. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–19.

[54] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144.

[55] Benjamin Shickel, Tyler J. Loftus, Lasith Adhikari, Tezcan Ozrazgat-Baslanti, Azra Bihorac, and Parisa Rashidi. 2019. DeepSOFA: A Continuous Acuity Score for Critically Ill Patients Using Clinically Interpretable Deep Learning. *Scientific Reports* 9, 1 (2019), 1879.

[56] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2017. Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. *IEEE journal of biomedical and health informatics* 22, 5 (2017), 1589–1604.

[57] Anima Singh, Girish Nadkarni, Omri Gottesman, Stephen B. Ellis, Erwin P. Bottinger, and John V. Guttag. 2015. Incorporating Temporal EHR Data in Predictive Models for Risk Stratification of Renal Function Deterioration. *Journal of Biomedical Informatics* 53 (2015), 220–228.

[58] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large Language Models Encode Clinical Knowledge. *Nature* 620, 7972 (2023).

[59] Venkatesh Sivaraman, Leigh A Bukowski, Joel Levin, Jeremy M. Kahn, and Adam Perer. 2023. Ignore, Trust, or Negotiate: Understanding Clinician Acceptance of AI-Based Treatment Recommendations in Health Care. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–18.

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

[61] Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *NPJ digital medicine* 3, 1 (2020), 17.

[62] Anja Thieme, Maryann Hanratty, Maria Lyons, Jorge Palacios, Rita Faia Marques, Cecily Morrison, and Gavin Doherty. 2023. Designing human-centered AI for mental health: Developing clinically relevant applications for online CBT treatment. *ACM Transactions on Computer-Human Interaction* 30, 2 (2023), 1–50.

[63] Anja Thieme, Abhijith Rajamohan, Benjamin Cooper, Heather Groombridge, Robert Simister, Barney Wong, Nicholas Woznitza, Mark A Pinnock, Maria T Wetscherek, Cecily Morrison, et al. 2025. Challenges for responsible AI design and workflow integration in healthcare: a case study of automatic feeding tube qualification in radiology. *ACM Transactions on Computer-Human Interaction* 32, 4 (2025), 1–61.

[64] Sana Tonekaboni, Shalmali Joshi, Melissa D McCradden, and Anna Goldenberg. 2019. What clinicians want: contextualizing explainable machine learning for clinical end use. In *Machine learning for healthcare conference*. PMLR, 359–380.

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

[66] Hein Minn Tun, Hanif Abdul Rahman, Lin Naing, and Owais Ahmed Malik. 2025. Trust in artificial intelligence–based clinical decision support systems among health care workers: Systematic review. *Journal of Medical Internet Research* 27 (2025), e69678.

[67] Alfredo Vellido. 2020. The Importance of Interpretability and Visualization in Machine Learning for Applications in Medicine and Health Care. *Neural Computing and Applications* 32, 24 (2020), 18069–18083.

[68] Colin G Walsh, Kavya Sharman, and George Hripcsak. 2017. Beyond discrimination: a comparison of calibration methods and clinical usefulness of predictive models of readmission risk. *Journal of biomedical informatics* 76 (2017), 9–18.

[69] 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.

[70] Xinru Wang and Ming Yin. 2021. Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 318–328.

[71] Jacob O Wobbrock, Leah Findlater, Darren Gergle, and James J Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 143–146.

[72] Sara Wolf, Tobias Grundgeiger, Raphael Zähringer, Lora Shishkova, Franzisca Maas, Christina Dilling, and Oliver Happel. 2025. How a Clinical Decision Support System Changed the Diagnosis Process: Insights from an Experimental Mixed-Method Study in a Full-Scale Anesthesiology Simulation. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–23.

[73] 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.

[74] Stephen B Woolley, Alex A Cardoni, and John W Goethe. 2009. Last-observation-carried-forward imputation method in clinical efficacy trials: review of 352 antidepressant studies. *Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy* 29, 12 (2009), 1408–1416.

[75] Xiao Xu, Zhiyuan Xu, Tiantian Ma, Shaomei Li, Huayi Pei, Jinghong Zhao, Ying Zhang, Zibo Xiong, Yumei Liao, Ying Li, et al. 2024. Machine learning for identification of short-term all-cause and cardiovascular deaths among patients undergoing peritoneal dialysis. *Clinical Kidney Journal* 17, 9 (2024), sfae242.

[76] Qian Yang, Aaron Steinfeld, Carolyn Rosé, and John Zimmerman. 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–13.

[77] 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.

[78] Nur Yildirim, Hannah Richardson, Maria Teodora Wetscherek, Junaid Bajwa, Joseph Jacob, Mark Ames Pinnock, Stephen Harris, Daniel Coelho De Castro, Shruthi Bannur, Stephanie Hyland, et al. 2024. Multimodal healthcare AI: identifying and designing clinically relevant vision-language applications for radiology. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–22.

[79] Evan You and Vue.js Team. 2020. Vue.js. https://github.com/vuejs/core

[80] Hubert D Zając, Tariq O Andersen, Elijah Kwasa, Ruth Wanjohi, Mary K Onyinkwa, Edward K Mwaniki, Samuel N Gitau, Shawnim S Yaseen, Jonathan F Carlsen, Marco Fraccaro, et al. 2025. Towards Clinically Useful AI: From Radiology Practices in Global South and North to Visions of AI Support. *ACM Transactions on Computer-Human Interaction* 32, 2 (2025), 1–38.

[81] Shao Zhang, Jianing Yu, Xuhai Xu, Changchang Yin, Yuxuan Lu, Bingsheng Yao, Melanie Tory, Lace M Padilla, Jeffrey Caterino, Ping Zhang, et al. 2024. Rethinking human-AI collaboration in complex medical decision making: a case study in sepsis diagnosis. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–18.

[82] Yinghao Zhu, Changyu Ren, Zixiang Wang, Xiaochen Zheng, Shiyun Xie, Junlan Feng, Xi Zhu, Zhoujun Li, Liantao Ma, and Chengwei Pan. 2024. EMERGE: Enhancing Multimodal Electronic Health Records Predictive Modeling with Retrieval-Augmented Generation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 3549–3559.

[83] Yinghao Zhu, Wenqing Wang, Junyi Gao, and Liantao Ma. 2024. PyEHR: A Predictive Modeling Toolkit for Electronic Health Records. https://github.com/yhzhu99/pyehr.

[84] Yinghao Zhu, Zixiang Wang, Long He, Shiyun Xie, Xiaochen Zheng, Liantao Ma, and Chengwei Pan. 2024. PRISM: Mitigating EHR Data Sparsity via Learning from Missing Feature Calibrated Prototype Patient Representations. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24)*. Association for Computing Machinery, New York, NY, USA, 3560–3569.

A Data Inclusion and Cohort Selection

This study utilizes retrospective datasets derived from the EHR data of three departments. To ensure data privacy and integrity, we de-identify all personally identifiable information (PII) prior to analysis.

A.1 Nephrology Cohorts Selection

We establish retrospective cohorts focusing on patients undergoing Peritoneal Dialysis (PD) from two medical centers: XY Hospital (covering December 2009 to October 2023) and BS Hospital (covering December 2011 to November 2021). To ensure the study population represents a stable adult PD cohort and to minimize confounding factors from other treatments or severe comorbidities, we apply identical inclusion and exclusion criteria across both institutions. Figure 5 illustrates the detailed participant selection process for both datasets.

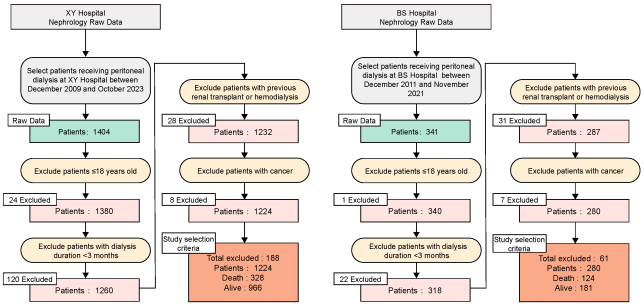

Figure 5: The flowchart of participant selection for nephrology cohorts. The left panel displays the screening process for XY Hospital, while the right panel depicts the process for BS Hospital. The procedure sequentially filters patients based on age, dialysis vintage, history of other renal therapies, and presence of malignant tumors.

Patients are included in the final cohorts if they meet the following criteria: (1) Adult status: The patient is older than 18 years at the time of data entry. This criterion excludes pediatric cases, which often require distinct clinical management strategies. (2) Stable dialysis vintage: The patient has maintained regular peritoneal dialysis for a duration of at least three months. This period allows for the stabilization of physiological parameters following the initial catheterization and dialysis induction.

To ensure the homogeneity of the dataset regarding the specific effects of peritoneal dialysis, we exclude patients based on the following conditions: (1) History of alternative renal replacement therapies: We exclude patients with a documented history of kidney transplantation or those who underwent emergency hemodialysis prior to PD. These treatments introduce physiological variations that differ significantly from those solely on peritoneal dialysis. (2) Malignant comorbidities: We exclude patients diagnosed with malignant tumors or cancer, as the systemic impact of malignancy and associated treatments (e.g., chemotherapy) significantly alters metabolic profiles and survival outcomes, potentially confounding the analysis of dialysis-related factors.

A.2 Obstetrics Cohorts Selection

We obtained obstetrics data from the EHR system of BC Hospital, covering the period from March 2003 to November 2024. Figure 6 illustrates the detailed participant selection process. In this study, we focused on the twin cohort.

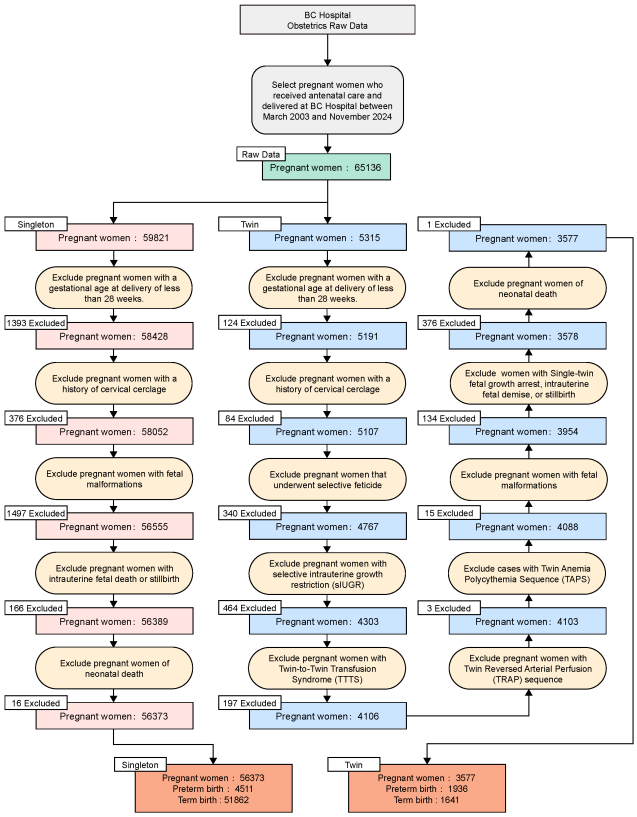

Figure 6: Participant selection flowchart for the obstetrics dataset. The workflow distinguishes between singleton and twin pregnancies, sequentially applying exclusion criteria regarding gestational age, medical history, and fetal complications.

We applied specific exclusion criteria to strictly define the study population and minimize confounding variables. For the singleton cohort, we excluded women based on the following conditions: (1) gestational age at delivery of less than 28 weeks, (2) a history of cervical cerclage, (3) fetal malformations, (4) intrauterine fetal death or stillbirth, and (5) neonatal death. For the twin cohort, the exclusion criteria further addressed complications specific to multiple gestations. In addition to gestational age less than 28 weeks and history of cervical cerclage, we excluded cases involving: (1) selective feticide, (2) selective intrauterine growth restriction (sIUGR), (3) twin-to-twin transfusion syndrome (TTTS), (4) twin anemia-polycythemia sequence (TAPS), (5) twin reversed arterial perfusion (TRAP) sequence, (6) fetal malformations, (7) single-twin fetal growth arrest, intrauterine fetal demise, or stillbirth, and (8) neonatal death. Following this selection process, the final dataset comprised 56,373 singleton pregnancies and 3,577 twin pregnancies.

B Dataset Features Statistics

This appendix presents the comprehensive statistics for the clinical features included in the three datasets used in this study, as detailed in Table 12 (XY Hospital), Table 13 (BS Hospital), and Table 14 (BC Hospital).

Table 12: Static and dynamic features statistics for the XY Hospital dataset ($N = 1,224$). Complete list of static ($N = 7$) and dynamic ($N = 33$) features for the XY Hospital dataset (Nephrology). Units for binary or categorical variables are denoted as '-'.

Feature	Unit	Feature	Unit	Feature	Unit
Static Features					
Height	cm	Chronic Nephritis/IgA	-	Polycystic Kidney	-
Weight	kg	Hypertensive Nephrop.	-	IgA Nephropathy	-
BMI	kg/m^2				
Dynamic Features					
Real-time Age	years	Sodium	mmol/L	Total Protein	g/L
Dialysis Vintage	years	Uric Acid	μmol/L	Total Cholesterol	mmol/L
Albumin	g/L	Glucose	mmol/L	TIBC	μmol/L
WBC	10^9/L	Prealbumin	mg/L	Resp. Sys. Diseases	count
ALT	U/L	AST	U/L	PD-related Comp.	count
LDL-Cholesterol	mmol/L	Ferritin	ng/mL	Cardio-cereb. Comp.	count
Calcium	mmol/L	Hemoglobin	g/L	Digestive Sys. Dis.	count
Triglycerides	mmol/L	iPTH	pg/mL	Acute Upper Resp.	count
HDL-Cholesterol	mmol/L	Platelets	10^9/L	PD-related Perit.	count
MCV	fL	ALP	U/L	Creatinine	μmol/L
Potassium	mmol/L	Phosphorus	mmol/L	Chloride	mmol/L

Table 13: Static and dynamic features statistics for the BS Hospital dataset ($N = 280$). Complete list of static ($N = 23$) and dynamic ($N = 66$) features for the BS Hospital dataset (Nephrology). Units for binary or categorical variables are denoted as '-'.

Feature	Unit	Feature	Unit	Feature	Unit
Static Features					
Gender	-	Pre-PD Albumin	g/L	Hypertensive Renal	-
Diabetes	-	Pre-PD Hemoglobin	g/L	Polycystic Kidney	-
BMI	kg/m^2	Pre-PD Creatinine	μmol/L	Obstructive Nephrop.	-
Hist. CHD	-	Diabetic Nephrop.	-	HSP Nephritis	-
Height (at Dialysis)	cm	Chronic Interstitial	-	IgA Nephropathy	-
Weight (at Dialysis)	kg	Chronic Glomerulo.	-	Ischemic Nephrop.	-
Hist. Old MI	-	Hist. Cerebral Infarc.	-	Hist. Cerebral Bleed	-
Follow-up Peritonitis	-	Pre-PD Emergency HD	-		
Dynamic Features					
Real-time Age	years	HCT	%	Neutrophil %	%
Dialysis Vintage	years	HDL-C	mmol/L	Phosphorus (P)	mmol/L
Albumin (ALB)	g/L	Hemoglobin (HGB)	g/L	P-LCR	%
ALP	U/L	Potassium (K)	mmol/L	Prealbumin (PAB)	mg/L
ALT	U/L	LDH	U/L	PCT	%
AST	U/L	LDL-C	mmol/L	PDW	fL
Basophil #	10^9/L	Lymphocyte #	10^9/L	Platelets (PLT)	10^9/L
Basophil %	%	Lymphocyte %	%	RBC	10^{12}/L
C1q	mg/L	MCH	pg	RDW-CV	%
CK	U/L	MCHC	g/L	RDW-SD	fL
CK-MB	U/L	MCV	fL	Total Bilirubin	μmol/L
Chloride (Cl)	mmol/L	Monocyte #	10^9/L	Total Cholesterol	mmol/L
Calcium (Ca)	mmol/L	Monocyte %	%	TCO2	mmol/L
Creatinine (Cr)	μmol/L	MPV	fL	Triglycerides (TG)	mmol/L
Direct Bilirubin	μmol/L	Sodium (Na)	mmol/L	Total Protein (TP)	g/L
Eosinophil #	10^9/L	Neutrophil #	10^9/L	Uric Acid (UA)	μmol/L
Eosinophil %	%	Urea	mmol/L	US-CRP	mg/L
Globulin (GLB)	g/L	WBC	10^9/L	Cystatin C	mg/L
HBDH	U/L	β2-MG	mg/L	γ-GT	U/L
MLR	ratio	PLR	ratio	MPV/PC Ratio	ratio
Resp. Sys. Diseases	count	Cardio-cereb. Dis.	count	Acute Upper Resp.	count
PD-related Comp.	count	Digestive Sys. Dis.	count	PD-related Perit.	count

Table 14: Static and dynamic features statistics for the BC Hospital dataset ($N = 3,577$, Twin Cohort). Complete list of static ($N = 29$) and dynamic ($N = 50$) features for the BC Hospital dataset (Obstetrics). Units for binary or categorical variables are denoted as '-'.

Feature	Unit	Feature	Unit	Feature	Unit
Static Features					
Maternal Age	years	Hist. C-Section	-	Preeclampsia	-
Height	cm	Family Hist. HTN	-	Preg. w/ Fibroids	-
Pre-pregnancy Weight	kg	Family Hist. Diabetes	-	Preg. w/ Adenomyosis	-
Hist. HTN	-	sIUGR	-	Chorionicity	-
Hist. Diabetes	-	GDM (Current)	-	PCOS	-
Hist. GDM	-	Pre-gestational Diab.	-	Nephrotic Syndrome	-
Hist. Preterm Birth	-	Gestational HTN	-	AKI	-
Hist. PROM	-	Pyelonephritis	-	SLE	-
Hist. Fibroids	-	OSA	-	APS	-
Recurrent Miscarriage	-	Sleep Breathing Dis.	-		
Dynamic Features					
Maternal Real-time Age	years	Total Protein (TP)	g/L	Creatinine (Cr)	μmol/L
Cervical Length	cm	Albumin (ALB)	g/L	Uric Acid (UA)	μmol/L
PT	s	Globulin (GLB)	g/L	CO2	mmol/L
APTT	s	Total Bilirubin (TB)	μmol/L	Sodium (Na)	mmol/L
Thrombin Time (TT)	s	Direct Bilirubin (DB)	μmol/L	Potassium (K)	mmol/L
Fibrinogen (Fib)	g/L	Total Bile Acid	μmol/L	Chloride (Cl)	mmol/L
D-Dimer	mg/L	Urea	mmol/L	Calcium (Ca)	mmol/L
APTT Ratio (PAPTT)	ratio	Phosphorus (P)	mmol/L	TSH	mIU/L
PT Ratio (PTT)	ratio	Fasting Plasma Glu	mmol/L	Thyroxine (T4)	μg/dL
ALT	U/L	Free Thyroxine (FT4)	ng/dL	Free T3 (FT3)	pmol/L
AST	U/L	WBC	10^9/L	Neutrophils (NE)	10^9/L
ALP	U/L	Monocytes (MO)	10^9/L	Basophils (BAS)	10^9/L
Eosinophils (EOS)	10^9/L	Neutrophil %	%	Monocyte %	%
Basophil %	%	Eosinophil %	%	RBC	10^{12}/L
Hemoglobin (Hb)	g/L	MCV	fL	RDW-CV	%
RDW-SD	fL	Platelets (PLT)	10^9/L	Mean Platelet Vol	fL
Plateletcrit (PCT)	%	PDW	fL		

C Detailed Data Preprocessing and EHR Model Training

This section provides a comprehensive description of the data processing, model architecture, training, and calibration procedures used to develop the predictive models for AICare. All procedures were implemented using our PyEHR toolkit [83].

C.1 Data preprocessing for nephrology.

The pipeline for the end-stage renal disease (ESRD) datasets (XY and BS Hospitals) transforms raw patient records into a structured format suitable for time-series modeling. The key steps include:

- **Data integration and cleaning.** Records from multiple sources were merged and patient identifiers were standardized. Temporal data, such as visit and birth dates, were validated and harmonized.
- **Feature engineering.** Static features (e.g., primary renal disease, height) and dynamic features (e.g., laboratory results, comorbidities) were separated. We computed Body Mass Index (BMI), patient age, and dialysis vintage (duration of treatment) at each clinical visit. Categorical features were one-hot encoded based on clinical keywords.
- **Cohort definition and labeling.** The predictive task was to forecast mortality within a 365-day window. To prevent immortal time bias, we excluded all records for surviving patients that occurred within 365 days of their final follow-up [57]. A binary label was then assigned to each visit: "1" for mortality within 365 days and "0" otherwise.
- **Imputation and normalization.** We used a 10-fold stratified cross-validation setup. To prevent data leakage, imputation and normalization statistics were derived solely from the training set

of each fold. Missing values were addressed first with a forward-fill imputation for each patient's timeline, followed by imputation of any remaining initial missing values using the training set median. All features were then Z-score normalized.
- **Class imbalance.** To mitigate the effects of an imbalanced class distribution, we applied random oversampling to the minority class within the training set of each fold.

C.2 Data preprocessing for obstetrics.

The pipeline for the obstetrics dataset (BC Hospital) predicting spontaneous preterm birth was largely consistent with the nephrology pipeline, with adaptations for the specific clinical context:
- **Data integration and identifier creation.** Data from outpatient, inpatient, laboratory, and demographic sources were integrated. A unique patient-pregnancy identifier was created to distinguish between multiple pregnancies for the same individual. Only records within 300 days prior to delivery were retained.
- **Cohort refinement.** A rigorous exclusion process was applied to create a homogeneous cohort. We removed cases with missing critical outcomes (e.g., gestational age at delivery) and those with clinical confounders such as cervical cerclage, selective fetal reduction, or fetal growth restriction.
- **Feature engineering and aggregation.** Static features (e.g., maternal age at delivery) and dynamic, time-varying features (e.g., gestational age at each visit) were created. To ensure a consistent time-series, multiple measurements within a single day were aggregated by their mean value. Features with a missingness rate above 90% were pruned.
- **Fold-specific preprocessing.** A 10-fold stratified cross-validation was performed at the patient level to ensure all records from a patient-pregnancy dyad remained in the same data split. Imputation and normalization followed the same leakage-prevention protocol as in the nephrology pipeline, using a forward-fill strategy followed by median imputation for dynamic features and Z-score normalization based on training set statistics.

C.3 Model architecture and training.

Our AICare model (adapted from ConCare), designed to model complex interactions in longitudinal EHR data [42]. The model processes patient trajectories through three main stages. First, a series of parallel Gated Recurrent Units (GRUs), one for each feature channel, captures the unique temporal evolution of each clinical variable. Second, these feature-specific representations are fed into a Transformer encoder block, where a multi-head self-attention mechanism models time-aware, cross-feature interactions. To encourage diverse representations, a decorrelation loss is applied to the attention head outputs. Finally, a terminal attention mechanism aggregates the contextualized feature vectors at each timestep to generate a dynamic patient representation and a corresponding risk prediction for each visit.

The model was implemented in PyTorch and trained using the PyTorch Lightning framework. We used the Adam optimizer with gradient clipping (max norm of 1.0) for numerical stability. The primary training objective used a Binary Cross-Entropy loss. An early stopping strategy was employed, monitoring the Area Under the Precision-Recall Curve (AUPRC) on the validation set. Training was halted if the validation AUPRC failed to improve for a specified number of epochs (patience), and the model checkpoint with the highest AUPRC was saved for evaluation.

C.4 Post-hoc calibration and thresholding.

To enhance the clinical utility of the model's predictions, we implemented a two-stage post-hoc optimization process on the validation set after training was complete [68].
- **Probability calibration.** We used Temperature Scaling to calibrate the model's outputs. This method learns a single scalar parameter (temperature, T) to rescale the pre-sigmoid logits, ensuring that the resulting probabilities more accurately reflect the true likelihood of the outcome. The optimal temperature is found by minimizing the Binary Cross-Entropy loss on the validation dataset [30].
- **Optimal decision threshold selection.** Following calibration, we determined the optimal decision threshold for converting probabilities into discrete class labels. We systematically searched through 200 candidate thresholds (from 0.01 to 0.99) to identify the value that maximized the F-beta score on the validation set. This allows the model's classification boundary to be tuned to the specific clinical costs of false positives versus false negatives.

C.5 Hyperparameter configuration.

The key hyperparameters used for training the AICare models on the three datasets are detailed in Table 15. These parameters were determined through a combination of standard practices and experimentation on the validation sets [38].

Table 15: Hyperparameters for the AICare models on the three study datasets. Final values were selected based on performance on the validation set. "Cls." means classification.

Hyperparameter	XY-Nephrology	BS-Nephrology	BC-Obstetrics
Task	Cls.	Cls.	Cls.
Seed	42	42	42
Epochs	30	30	30
Patience (early stopping)	10	10	5
Batch size	32	16	128
Learning rate	0.001	0.001	0.01
Primary metric	AUPRC	AUPRC	AUPRC
Hidden dimension	128	128	32
Output dimension	1	1	1
Dynamic feature dimension	33	66	50
Static feature dimension	7	23	29

D Prompt for Generating Diagnostic Recommendations

```
System prompt:
You are an experienced clinician with extensive medical knowledge and clinical diagnostic experience.

You will receive a patient's electronic health record (EHR) data, an AI model's risk prediction result, and feature importance weights. Based on this information, please conduct a clinical analysis and provide diagnostic and decision-making recommendations.

Analysis Requirements:
1. Focus on the examination values from the most recent visit and the features with high importance weights.
2. Use analytical reasoning to deduce the patient's physiological or biochemical pathophysiological state.
3. Systematically identify the appropriate clinical response.
4. Provide specific clinical advice without listing the patient's specific data (do not use concrete numerical values).
5. Ensure the response is detailed and substantial.


User prompt template:
**Clinical Prediction Task**
{task_definition}

**Patient's Electronic Health Record Analysis**
AI Model Risk Prediction Result: {risk_value}%

Feature Importance Weights (Key Factors Influencing Prediction):
{key_features_list}

Patient's Complete Examination Values from the Last Visit:
{all_feature_values}

**Clinical Analysis Request**
Based on the AI model's analysis results and the patient's EHR data above, please use clinical reasoning to analyze the patient's pathophysiological state and provide specific diagnostic and decision-making recommendations.
```

To generate high-quality, personalized clinical advice, we have carefully engineered the prompt for interacting with a large language model (LLM). The prompt is constructed from two parts: a static system prompt that defines the model's role, core objectives, and output constraints, and a dynamic user prompt template that structures the specific patient data and AI analysis results for the task at hand. This separation ensures consistent model behavior while allowing for flexible adaptation to different clinical tasks.

E Guidance Script for Participants

This script provides a consistent introduction for researchers to guide clinician participants through the study's questionnaires and interviews.

Introduction. "Thank you once again for your invaluable contribution to our research on the AICare system. Your professional insights are essential for us to understand the practical implications and potential of this technology. Next, we will conduct experiments, including a few standardized questionnaires and a brief concluding interview. Your honest and detailed feedback is deeply appreciated."

Part 1: Participant background questionnaire. "First, we have a brief background questionnaire to understand your professional profile, including your specialty, clinical experience, and general familiarity with AI technology. This information is purely for data analysis and helps us interpret the results across different clinician groups. There are no right or wrong answers."

Part 2: Standardized assessment scales. "In this section, we will use three internationally recognized assessment scales. These instruments are standard in human-computer interaction research and allow us to quantitatively measure key aspects of your experience, such as perceived workload, system usability, and trust. Using these validated scales ensures our findings are objective and comparable to established benchmarks in the field."

"I will briefly explain each scale:"

(1) "**NASA Task Load Index (NASA-TLX): Measuring your cognitive workload.** This scale assesses the mental and physical effort required to complete the diagnostic tasks. It is a two-step process. First, for six dimensions like 'Mental Demand' or 'Temporal Demand,' you will mark a line to indicate your feeling, from low to high. Please note that you will complete this questionnaire twice: once after performing the diagnostic tasks without assistance, and again after using the AICare system."

(2) "**System Usability Scale (SUS): Measuring how easy the system is to use.** This scale directly measures whether you found the AICare system to be straightforward, convenient, and well-designed. There are 10 statements; please indicate your level of agreement with each. Note that some statements are phrased positively (e.g., 'I thought the system was easy to use') and others are phrased negatively (e.g., 'I found the system unnecessarily complex'). Please read each statement carefully."

(3) "**Trust in Automation Scale: Measuring your trust in the AI.** This scale assesses your confidence in the analysis and recommendations provided by AICare. Given that trust is the foundation of human-AI collaboration in medicine, your perspective here is vital. Similar to the SUS, please read each statement and choose the option on the 7-point scale that best represents your opinion."

"Finally, before we move to the interview, we want to distinguish between the system as a whole and its specific parts. Listed here are the four main functional modules of AICare (the risk trajectory, the factor list, the LLM recommendation, and the population-level analysis). Please rate how useful you found each specific feature for your diagnostic process on a scale of 1 to 5."

Part 3: Semi-structured interview. "To conclude our session, we will have a semi-structured interview lasting approximately 15 minutes. While the questionnaires provide valuable quantitative data, this conversation is our opportunity to understand the context and the 'why' behind your ratings. We are especially interested in your thought processes, specific experiences with the system, and your valuable clinical perspective."

F Study Instruments

This appendix contains all questionnaires and the interview guide used in the study.

F.1 Participant Background Questionnaire

Please select or fill in the option that best describes you.

(1) **Participant ID:** _____
(2) **Name:** _____
(3) **Gender:** ◯ Male ◯ Female ◯ Other / Prefer not to say
(4) **Age range:** ◯ <30 ◯ 30-39 ◯ 40-49 ◯ 50+
(5) **What is your clinical specialty?**
- ◯ A. Nephrology
- ◯ B. Obstetrics
- ◯ C. Other (Please specify): _____

(6) **How many years have you been in clinical practice?**
- ◯ A. Less than 1 year or in internship/rotation
- ◯ B. 1-5 years
- ◯ C. 6-10 years
- ◯ D. 11-15 years
- ◯ E. 16-20 years
- ◯ F. More than 20 years

(7) **What is your current role or professional title?**
- ◯ A. Medical Student (currently enrolled)
- ◯ B. Intern / Resident in Training
- ◯ C. Resident Physician
- ◯ D. Attending Physician
- ◯ E. Associate Chief Physician
- ◯ F. Chief Physician
- ◯ G. Other (Please specify): _____

(8) **How familiar are you with AI-powered clinical decision support systems?**
- ◯ 1: Not familiar at all (I have not heard of this concept).
- ◯ 2: Slightly familiar (I have heard of it, but am not sure what it does).
- ◯ 3: Moderately familiar (I understand its general definition and purpose).
- ◯ 4: Very familiar (I have a deeper understanding of its applications and key technologies).
- ◯ 5: Expert (I have professional knowledge or practical experience in this field).

(9) **Prior to this study, how often did you use AI-based decision support systems in your clinical work?**
- ◯ A. Never
- ◯ B. Rarely (e.g., once or twice a month)
- ◯ C. Occasionally (e.g., a few times a month)
- ◯ D. Frequently (e.g., on most workdays)

F.2 Case-specific Questionnaire

For the patient case you just reviewed, please answer the following questions.

Part 1: Nephrology Cases

(1) **What is your assessment of this patient's risk of mortality within the next year?**
- ◯ A. 0-25% (Low risk): Very unlikely to die within 1 year.
- ◯ B. 25-50% (Low-medium risk): Generally stable, but with some health risks.
- ◯ C. 50-75% (Medium-high risk): Significant health concerns that require attention.
- ◯ D. 75-100% (High risk): Critical health indicators, high probability of death within 1 year.

(2) **How confident are you in this assessment?** (1=Not confident at all, 5=Very confident)
◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5

(3) **Which features were most important for your assessment? (Select all that apply)**
□ Albumin □ Diastolic BP □ Systolic BP □ Urea □ Weight
□ Serum Sodium □ Serum Chloride □ Hemoglobin □ Creatinine
□ Serum Phosphorus □ Serum Potassium □ Food Intake □ Serum Calcium
□ Bicarbonate □ hs-CRP □ Blood Glucose □ White Blood Cell Count
□ Other: _____

Part 2: Obstetrics Cases

(4) **What is your assessment of this patient's risk of spontaneous preterm birth (before 37 weeks)?**
- ◯ A. 0-25% (Low risk): Very likely to carry to term.
- ◯ B. 25-50% (Low-medium risk): Some risk factors present, but preterm birth is not expected.
- ◯ C. 50-75% (Medium-high risk): Significant concern for preterm birth, may require intervention.
- ◯ D. 75-100% (High risk): Preterm birth is highly probable without intervention.

(5) **How confident are you in this assessment?** (1=Not confident at all, 5=Very confident)
◯ 1 ◯ 2 ◯ 3 ◯ 4 ◯ 5

(6) **Which features were most important for your assessment? (Select all that apply)**
□ Activated Partial Thromboplastin Time (APTT)
□ Prothrombin Time (PT) □ Thrombin Time (TT)
□ Fibrinogen (Fib)
□ Thyroid Stimulating Hormone (TSH) □ Thyroxine (T4)
□ Albumin (ALB) □ Total Protein (TP)
□ Potassium (K) □ Chloride (Cl) □ Glucose (GLU)
□ Red Blood Cell Count (RBC) □ Other: _____

F.3 NASA-Task Load Index (NASA-TLX)

This scale assesses the workload you experienced during the last block of tasks. Please place a mark on each scale that matches your experience.

- **Mental Demand**: How much mental and perceptual activity was required (e.g., thinking, deciding, remembering, connecting patient data)?
  Low _____ High
- **Physical Demand**: How much physical activity was required (e.g., clicking, scrolling, interacting with the interface)?
  Low _____ High
- **Temporal Demand**: How much time pressure did you feel due to the rate or pace at which the tasks needed to be completed?
  Low _____ High

- **Performance**: How successful do you think you were in accomplishing the goals of the diagnostic task? (Note the direction).
  Good _____ Poor
- **Effort**: How hard did you have to work (mentally and physically) to accomplish your level of performance?
  Low _____ High
- **Frustration**: How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, and relaxed did you feel during the task?
  Low _____ High

F.4 System Usability Scale (SUS)

The specific items used in the System Usability Scale are listed in Table 16.

F.5 Trust in Automation Scale

The items comprising the Trust in Automation Scale are provided in Table 17.

F.6 AICare Feature Feedback

Please rate the usefulness of each AICare feature for your diagnostic process. (1=Not useful at all, 5=Extremely useful). The feedback form structure is shown in Table 18.

F.7 Semi-structured Interview Guide

Researcher's Guide: The goal is to understand the "why" behind the quantitative results. Use these questions as a guide, but feel free to ask follow-up questions to probe deeper into interesting comments. Start by making the participant feel comfortable.

(1) **On Usability and Workflow Integration:**
   - How easy or difficult was it to understand the information presented by AICare? Was any part of the interface confusing or overwhelming?
   - Imagine if your department introduced this tool tomorrow. How do you think it would integrate into your daily work?
   - Do you think it could help you improve diagnostic efficiency and accuracy?

(2) **On Trust, Interpretability, and Clinical Reasoning:**
   - Let's talk about trust. When you were using AICare, were there any specific moments or features that made you start to trust its predictions? Conversely, were there times you felt skeptical?
   - I would like to talk specifically about a few of the explanatory features:
     – How did the **dynamic risk trajectory visualization** influence your understanding of the patient's condition? Did seeing the historical and predicted trends help you build a holistic view of the patient's situation?
     – What about the **interactive list of critical risk factors**? Can you recall an instance where you hovered over or clicked on an indicator to see its trend? How did that action affect your trust in the AI's reasoning?
     – Finally, what are your thoughts on the **LLM-driven diagnostic recommendation**? Did you find them useful? Did they feel consistent with the other explanatory features?
     – The system also showed **population-level indicator analysis visualization**. Did you find this feature to be a useful supplement to the AI's explanation, or was it redundant?
   - During your use, was there a time when your own clinical judgment and the AI's prediction were inconsistent? Can you describe that situation and how you handled the conflict? Did the system's explanatory features help you understand the difference?

(3) **On Clinical Value and Impact:**
   - Did the system ever provide you with a new insight or make you notice a risk factor you might have otherwise overlooked? Can you give an example?

(4) **Closing Questions:**
   - Is there anything else you would like to share about your experience today?

Table 16: The system usability scale (SUS). For the following statements about the AICare system, please indicate your level of agreement.

#	Statement	Strongly Disagree (1)	Disagree (2)	Neutral (3)	Agree (4)	Strongly Agree (5)
1	I think that I would like to use this system frequently in my clinical work.	○	○	○	○	○
2	I found the system unnecessarily complex.	○	○	○	○	○
3	I thought the system was easy to use.	○	○	○	○	○
4	I think I would need the support of a technical person to be able to use this system.	○	○	○	○	○
5	I found the various functions in this system (e.g., charts, lists) were well integrated.	○	○	○	○	○
6	I thought there was too much inconsistency in this system.	○	○	○	○	○
7	I would imagine that most clinicians would learn to use this system very quickly.	○	○	○	○	○
8	I found the system very cumbersome to use.	○	○	○	○	○
9	I felt very confident using the system.	○	○	○	○	○
10	I needed to learn a lot of things before I could get going with this system.	○	○	○	○	○

Table 17: The trust in automation scale. Please indicate your agreement with the following statements about the AICare system. (1=Strongly Disagree, 7=Strongly Agree). (R) indicates a reverse-scored item.

#	Statement	1	2	3	4	5	6	7
1	The system is deceptive (e.g., it seems to hide important information or provide misleading results). (R)	○	○	○	○	○	○	○
2	The system's behavior is predictable (i.e., its analysis is consistent for similar patient cases).	○	○	○	○	○	○	○
3	The system is designed with the patient's best interest in mind.	○	○	○	○	○	○	○
4	I am unsure about the system's capabilities (i.e., when it is reliable and when it might fail). (R)	○	○	○	○	○	○	○
5	I am confident in the system.	○	○	○	○	○	○	○
6	The system has characteristics that could be harmful to my clinical decision-making. (R)	○	○	○	○	○	○	○
7	I am familiar with how the system works.	○	○	○	○	○	○	○
8	I can trust the system.	○	○	○	○	○	○	○
9	I am unfamiliar with how the system works. (R)	○	○	○	○	○	○	○
10	The system's explanations are a faithful representation of its reasoning (i.e., it honestly reflects its underlying data and logic).	○	○	○	○	○	○	○
11	I distrust the system. (R)	○	○	○	○	○	○	○
12	The system is reliable (i.e., its performance is stable and its findings are reproducible).	○	○	○	○	○	○	○

Table 18: Participant rating of the usefulness of each AICare feature.

Feature	1	2	3	4	5
Dynamic risk trajectory visualization	○	○	○	○	○
Interactive list of critical risk factors	○	○	○	○	○
Population-level indicator analysis visualization	○	○	○	○	○
LLM-driven diagnostic recommendation	○	○	○	○	○