The Causal Artificial Intelligence Clinician for early haemodynamic management of septic shock in ICU

Giovanni Angelotti[a,b], Laura Azzimonti[a,b,*], Maurizio Cecconi[c,d,1], Marco Zaffalon[a,1]

[a]*SUPSI, Dalle Molle Institute for Artificial Intelligence, Lugano, Switzerland*
[b]*Swiss Institute of Bioinformatics, Lugano, Switzerland*
[c]*Humanitas University, Department of Biomedical Sciences, Rozzano, Milan, Italy*
[d]*IRCCS Humanitas Research Hospital, Department of Anaesthesia and Intensive Care, Rozzano, Milan, Italy*

Abstract

Background: Standardizing fluid and vasopressor resuscitation in septic shock is challenging due to patient heterogeneity. We trained a causal model to identify optimal dosing during the first six hours of intensive care unit (ICU) admission.

Methods: Graphical causal inference models were applied to estimate heterogeneous treatment effects. Grounding models in expert clinical knowledge minimizes bias from spurious correlations to generate robust, contextually meaningful recommendations. Our model was trained on 1,702 MIMIC database admissions and externally validated on 1,434 eICU admissions. Primary outcomes were in-hospital survival and 24-hour clinical improvement (SOFA score reduction of two points or more).

Findings: The cohort comprised 3,136 participants (median age 65 years

*Corresponding author
[1]These authors contributed equally to this work.

[IQR 53–75]; 42.7% female). Deviation from vasopressor recommendations was associated with increased in-hospital mortality (median OR 5.61, 95% CI 5.44–5.78) and failed clinical improvement (median OR 6.33, 95% CI 6.17–6.50). Fluid deviations yielded corresponding median ORs of 1.02 (95% CI 1.02–1.02) and 1.14 (95% CI 1.14–1.14). In external validation, the model achieved a median survival AUROC of 0.73 (95% CI 0.69–0.77) and clinical improvement AUROC of 0.69 (95% CI 0.66–0.72), matching predictive baselines. Treatment effects were heterogeneous: optimal fluids increased survival by up to 4% in low-severity subgroups, while vasopressor responses varied from 0.5% to 17% across acute severity levels. Sensitivity analyses across 36 scenarios confirmed primary associations in 33 cases (91.7%).

Interpretation: Recommendations from expert-grounded causal models correlate with improved septic shock outcomes in external validation, capturing significant heterogeneity in patient response.

Funding: None

Research In Context

Evidence before this study

Early haemodynamic management of septic shock, involving fluid resuscitation and vasopressor administration, is critical to prevent organ failure and improve survival. Standardized protocols, such as those from the Surviving Sepsis Campaign, offer general guidance but may not fully account for the high clinical heterogeneity of patients. We searched PubMed for articles published up to July 2026 using the query: ("deep learning" OR

"reinforcement learning" OR "machine learning" OR "artificial intelligence" OR "causal inference" OR "causal AI" OR "target trial*") AND ("sepsis" OR "septic shock") AND ("fluid*" OR "vasopressor*" OR "resuscitation"). Among the 219 relevant studies identified, 141 proposed a model. Of these, the majority (120) relied on associative statistical relationships, including supervised (90), unsupervised (16), or reinforcement (14) learning frameworks, which can generate recommendations susceptible to confounding and spurious correlations. The remaining models based on causal inference (21) primarily utilized potential outcomes (17) or target trial emulation methods (4); while methodologically rigorous, these approaches do not explicitly encode clinical and physiological knowledge into their design. Incorporating such domain expertise constrains the model's parameter space to physiologically plausible relationships, generating treatment recommendations that are fundamentally aligned with established clinical workflows. In this study, we address this limitation by using structural causal models to integrate granular, physiology-grounded clinical reasoning into artificial intelligence models, supporting early haemodynamic management in septic shock.

Added value of this study

This study introduces the Causal Artificial Intelligence Clinician, a framework that utilizes Structural Causal Models (SCMs) to estimate the heterogeneous treatment effects of fluids and vasopressors during the first six hours of ICU admission. By integrating expert clinical consensus directly into the causal graph, the model controls for confounding using the backdoor criterion, ensuring that recommendations are grounded in physiological principles rather than simple association. Trained on 1,702 admissions from

the MIMIC-IV database and externally validated on 1,434 admissions from the eICU database, patients whose administered treatments aligned with the model's recommendations showed higher rates of in-hospital survival and clinical improvement. Crucially, the causal model achieved predictive performance comparable to conventional predictive baselines while utilizing approximately 65% fewer variables, substantially reducing data requirements and computational complexity.

Implications of all the available evidence

Our findings suggest that expert-grounded causal AI can provide robust and transparent decision support for complex, heterogeneous conditions like septic shock. Because these models are leaner and require fewer variables than standard predictive systems, they are potentially easier to implement, scale, and deploy in diverse clinical settings, including low- and middle-income countries with limited digital infrastructure. Furthermore, the explicit nature of the causal graph offers a transparent, shared representation that clinicians can scrutinize and refine, fostering clinical trust while ensuring that healthcare professionals retain final decision-making authority.

Introduction

The early hours for septic shock patients, which include the historically termed *golden hour*, represent a period of rapidly evolving tissue hypoperfusion and cytopathic hypoxia where delayed intervention causes irreversible organ dysfunction.[1,2,3] Current clinical guidelines advocate for immediate fluid resuscitation and early vasopressor initiation to restore mean arterial pressure and optimize oxygen delivery.[4,5,6] However, host-response heterogeneity fre-

quently renders standardized protocols insufficient.[7,8] Evidence suggests that tailoring the dosing of vasopressors and fluids in these first hours may play a key role in mitigating immediate haemodynamic instability and ultimately increasing the chances of survival.[9] Determining the optimal treatment combination remains, however, challenging. The design of Randomized Controlled Trials (RCTs) is complicated by the extreme heterogeneity of sepsis, the lack of known biomarkers, and poorly understood dosing thresholds.[10,11] Concurrently, intensive care units are among the most data-rich environments across clinical practice, offering vast repositories of observational data.[12,13] Consequently, estimating the effect of resuscitation strategies in sepsis using these observational records has been a recurrent objective of modern artificial intelligence systems.[14,15] These applications rely on predictive models that exploit statistical correlations to generate risk estimates for outcomes. However, they do not assist clinicians in identifying optimal treatment strategies to transition patient states toward improved clinical conditions.[16] There is therefore a growing need for AI systems capable of operationalizing evidence as a bridge that *concretely* connects actions to patients outcomes.[17] Causal inference offers a principled response to this need, providing the methodological tools to move beyond association and formally reason about the effects of interventions.[18] These methods have an established history in healthcare, now gaining renewed momentum through the scalability of novel Machine Learning (ML) models,[19,20] known as Causal AI.[21]

This manuscript adopts Structural Causal Models (SCMs), which represent causal relations as arcs in a graph.[22] Predictions made by models built in this manner are guaranteed to be consistent with the encoded assumptions,

thereby enabling the estimation of treatment effects under the appropriate conditions.[23]

In intensive care settings, SCMs have been applied to model oxygen therapy effects on mortality,[24] health equity disparities,[25] discharge decisions,[26] and have been used to audit sepsis treatments models.[27] Here, we introduce the *Causal Artificial Intelligence Clinician*, a framework that leverages SCMs to identify the optimal resuscitation strategy in the first six hours of intensive care for septic shock patients, formally encoding the clinical, physiological, and operational knowledge of healthcare professionals into a Causal AI model. We demonstrated in external validation that patients whose physicians administered treatments aligned with model suggestions had, on average, better outcomes in terms of in-hospital survival and clinical improvement measured as increased haemodynamic stability. The proposed framework remains robust to distribution shifts when applied to data from hospitals not used in training, while relying on fewer, more causally meaningful variables than predictive baselines. Together, these results show how encoding domain knowledge into the design of AI models enables robust estimation of treatment effects from the secondary use of electronic health records.

This analysis is structured into three main phases, as illustrated in Figure 1. First, an initial deliberation phase is conducted to identify a set of clinically robust assumptions (panel (a)). Then, the assumptions encoded in a causal graph guide the training of ML models (panel (b)). Finally, the resulting models are evaluated on an external validation set (panel (c)). TRIPOD+AI and Target Statement checklists for this study are reported

in Supplementary D. Our work is fully reproducible, instructions to build the datasets and re-run our analysis is available at: `https://github.com/IDSIA/causal-ai-clinician`.

Methods

Assumptions and Causal Implications

This study employs SCMs to estimate the effect of intravenous fluid and vasopressor doses on patient outcomes from observational data, distinguishing causal effects from spurious statistical correlations by incorporating clinical evidence and domain knowledge into the model structure through a causal graph. This creates a framework for simulated interventions that quantify survival and improvement probabilities under specific treatment scenarios, providing a formal approximation of outcomes traditionally observed in RCTs.

Specifically, for each patient, the model estimates the expected change in the probability of in-hospital survival or clinical improvement when administering specific doses of norepinephrine and intravenous fluids compared to no treatment. This can be formalized as a *causal contrast*, here the Conditional Average Treatment Effect (CATE), as

$$\text{CATE}((n, f), \mathbf{X}) = P(Y|do(N = n), do(F = f), \mathbf{X}) - P(Y|do(N = 0), do(F = 0), \mathbf{X})$$

where $do(N = n)$ and $do(F = f)$ represent *do* operators, corresponding to the act of intervening on a set of variables, in our case intravenous fluid dose F and norepinephrine N, forcing them to take specific values n and f respectively, regardless of their causal parents. Accordingly, $P(Y|do(N = 0), do(F = 0), \mathbf{X})$ represents the probability of outcome Y (where $Y = 1$ in

our case is either in-hospital survival or clinical improvement as opposed to $Y = 0$) following intervention on F and N in a population with baseline clinical profile $\mathbf{X}$.

Note that computing CATE via the simple observational probability $P(Y|N = n, F = f, \mathbf{X})$ is generally a mistake, as it is possible only under strict assumptions of unconditional unconfoundedness, which is not applicable in our case. We instead exploit the clinical knowledge encoded in the SCM to robustly measure the CATE from the data. We employ the *Backdoor Criterion* to identify which features should or should not be included in our model in order to properly recover the treatment effect. The Backdoor criterion can be verified from a causal graph alone and guarantees that if we can adjust for all parents $\mathbf{Z}$ of F and N the treatment effect can be recovered from the data[28](for a rigorous formulation of the Backdoor definition, please refer to corollary 3 in[22]). If the criterion holds, we can convert the causal contrast into a statistical estimator solvable with common machine learning models:

$$P(Y|do(N = n), do(F = f), \mathbf{W}) = \sum_{\mathbf{z}} P(Y|N = n, F = f, \mathbf{Z} = \mathbf{z}, \mathbf{W})P(\mathbf{Z} = \mathbf{z}|\mathbf{W})$$

where $\mathbf{W}$ are clinical descriptors one may be interested in conditioning for. Here we estimate the CATE using gradient boosting trees.Details are provided in Supplementary C.

Causal Graph Derivation

The causal graph was developed through iterative multidisciplinary consensus among clinicians and causal methodologists, encoding clinical assumptions about early haemodynamic management in septic shock. While the graph necessarily constitutes a simplification of reality, it provides a structural rationale grounded in clinical decision making. It incorporates both

observed variables, retrieved from MIMIC and eICU, and unmeasured confounders not available in the Electronic Health Record (EHR). Further details on causal graph derivation are provided in Supplementary C. A predictive baseline model was constructed for benchmarking purposes using a broader set of clinical and administrative markers, detailed in Table S1 in Supplementary B.

Cohort Derivation

Data for this study were extracted from the MIMIC-IV v3.1 Database (MIMIC) and eICU v2.0 Collaborative Research Database (eICU). The MIMIC database contains EHR from ICU patients collected at the Beth Israel Deaconess Medical Center in Boston, US, from 2008 to 2022.[12] The eICU is a multi-center EHR database with over 200,000 ICU admissions across the United States between 2014 and 2015.[13] To isolate the septic patients of interest, we first restricted our cohort to all patients admitted to the hospital from the Emergency Department (ED). To ensure we captured the earliest available septic shock dynamics, we included only ICU admissions (T_{0H}) occurring within 6 hours of the earliest available patient-hospital interaction ($T_{\text{-}6H}$), defined as ED admission in MIMIC and hospital admission in eICU. We further restricted the minimum ICU stay length to 30 hours (T_{30H}), the shortest time interval over which we could compute a complete updated SOFA score. This was necessary to compute our endpoint of clinical improvement without being influenced by measurements and interventions occurring during the first 6 hours of the ICU stay [T_{0H}, T_{6H}] , which defines the window used for feature extraction. We excluded patients admitted for surgery to avoid confounding due to trauma and patients with a Do Not Resuscitate

(DNR) order during the first 30 hours of their ICU stay to ensure continuity of care. In eICU, we additionally excluded patients who were subsequently transferred to avoid survival bias induced by severe cases being transferred to other centers. Variables extracted from the first 6 hours of the ICU stay [T_{0H},T_{6H}], were also used to filter our cohort. To reduce the likelihood of false-positive inclusions, we removed all patients with approximate lactate levels lower than 2 mmol/L and no evidence of either fluid therapy, antibiotic therapy, or a mean blood pressure lower than 65 mmHg without evidence of vasopressor therapy. For vasopressors, we considered dopamine, epinephrine, norepinephrine, phenylephrine, vasopressin, dobutamine, and milrinone. Details regarding the study design are shown in Figure S1 in Supplementary B.

Data Preparation

Given the complexity of navigating MIMIC and eICU, we relied as much as possible on publicly available, validated queries for the data extraction pipeline. The Glasgow Coma Scale defaulted to 15 for sedated patients. All features were extracted and aggregated from official materialized views within the [T_{0H}, T_{6H}] window; fluids represented cumulative intake, and norepinephrine denoted average doses, with undocumented treatments assumed as zero. Study outcomes included EHR-recorded in-hospital survival and clinical improvement, defined as a minimum two-point SOFA score reduction between the initial [T_{0H},T_{6H}] and subsequent [T_{30H},T_{54H}] intervals. Clinical outcomes were defined as follows: in-hospital survival was extracted directly from the EHR, while clinical improvement was defined as a reduction of at least 2 points between the worst SOFA scores recorded in the [T_{30H},T_{54H}]

and [T_{0H},T_{6H}] time-intervals. Details on features and their aggregation are shown in Table S1 in Supplementary B.

Model Training and Predictive Performance Assessment

We trained the causal model on the MIMIC dataset and externally validated it using eICU. Causal models included as input features as sufficient set as defined by our causal graph, as listed in Figure 2, utilizing LightGBM gradient boosted trees. We used doubly nested cross-validation, applying an inner loop for hyperparameter tuning and an outer loop for calibration to accurately estimate treatment effects. Overall performance and fairness were evaluated via Area Under the Receiver Operating Characteristic curve (AUROC) over 2,000 bootstraps, stratifying by gender and race to assess parity against male and White majority reference groups. Further details are provided in Supplementary C.

Optimal Intervention Estimation and Validation

For each patient, the optimal combination of vasopressor and fluid doses was identified as the one maximizing the estimated CATE, evaluated across a discrete grid of clinically observed dose ranges. To validate the estimated policies, patient outcomes in the external validation set were compared according to the deviation between administered treatments and model recommendations, under the assumption that alignment with the model correlates with improved patient prognosis. The association between treatment deviation and patient outcomes was quantified using repeated cross-validation and bootstrapping. Treatment effect heterogeneity across patient subgroups was further assessed by estimating CATEs conditioned on different baseline clinical profiles, with profiles clustered by common treatment-response trends

to identify clinically meaningful examples. Further details are provided in Supplementary C.

Sensitivity Analysis

Finally, we performed a series of sensitivity analysis to test the robustness of our design choices spanning different inclusion and exclusion criteria and modeling choices. To test robustness against potential dilution due to false-positive inclusions, particularly in eICU, we restricted our cohort to patients receiving vasopressor therapy, regardless of their blood pressure readings. Similarly, in one analysis we restricted inclusions to only those patients with explicit mentions of sepsis in the admission notes. We then investigated the impact of our Missing at Random (MAR) assumption for lactates by selecting only patients with lactate readings strictly above 2 mmol/L, excluding those with missing ones. In two additional sensitivity analyses we examined how including demographic and operational variables as features in our model, namely gender, hypertension and vasopressor status at the first available blood pressure reading, affects performance. Lastly, we assessed the effect of including DNR patients. We also evaluated the implications of using simpler linear models, a logistic regression, as a less overfitting-prone alternative. In the end, we investigated the effect of different censoring windows, 24 and 18 hours as against 30, to potentially identify effects linked to survivorship bias.

Results

Study Design and Cohort Characterization

The study included 3,136 eligible admissions, utilizing MIMIC-IV for training and eICU for external validation. We evaluated two primary end-points: clinical improvement, defined as a minimum two-point reduction in

the SOFA within the subsequent 24 to 48 hours, and overall in-hospital survival. These endpoints were observed in 12.2% and 80.8% of admissions, respectively. The analyzed treatments included total intravenous fluid intake and average norepinephrine equivalent dose during this initial window. See tables S1, S2 and S3 in Supplementary B for more details.

Causal graph derivation and physiological grounding

We reconstructed the clinical reasoning process underlying the physician's decision-making. At its core, the physician acts on two therapeutic interventions, fluid administration and norepinephrine, both aimed at modulating the downstream haemodynamic state: fluids influence cardiac output through preload optimization, while norepinephrine acts on systemic vascular resistance. We further assume that physicians rely on a set of specific physiological triggers to reach their treatment decisions. Fluid titration is driven by markers of hypoperfusion and renal stress, namely lactate, mean arterial pressure, and urine output. Norepinephrine doses, instead, are adjusted in response to diastolic pressure and heart rate, which themselves reflect the current degree of organ failure. Underlying this entire process is the complex interaction between the host's response to infection, for which we assume that an appropriate antibiotic was selected earlier in the emergency room, and the patient's age, which guides dose adjustment. Taken together, these assumptions are encoded in the causal graph shown in Figure 2. This graph allows us to translate our clinical question, namely:

"What is the probability of survival or clinical improvement if we administer a different dosage combination of norepinephrine and intravenous fluid in the early hours following ICU admission in septic shock patients?",

into a well-defined and rigorous causal contrast.

Optimal treatments, confounding and clinical outcomes

In Structural Causal Model (SCM) semantics, simulating an intervention involves removing incoming arcs to treatment nodes (Figure 2) and fixing them to target values, mimicking a Randomized Controlled Trial (RCT). The CATE is estimable via machine learning conditioned on the parent nodes of the treatment nodes, which form a sufficient adjustment set despite latent confounding from unrecorded infection and antibiotic responses.[29,30] Excessive vasopressor and fluid administration relative to the optimal causal model policy correlates with lower in-hospital survival (Figure 4a), while restrictive fluid policy correlates with poorer clinical improvement (Figure 4b). Every unit of vasopressor deviation reduces the odds of survival and clinical improvement by a median of 5.61 (95% CI: [5.44, 5.78]) and 6.33 (95% CI: [6.17, 6.50]), respectively. Fluid deviations reduce these odds by 1.02 (95% CI: [1.02, 1.02]) and 1.14 (95% CI: [1.14, 1.14]) (Table S8). The dose-response space mapping is highly non-linear (Figure 4). The causal model demonstrates predictive robustness. In external validation, it achieves median AUROCs of 0.73 (95% CI: [0.69, 0.77]) for survival and 0.69 (95% CI: [0.66, 0.72]) for clinical improvement. This matches a gradient boosting baseline trained on a larger superset (Table S1), which yields AUROCs of 0.74 (95% CI: [0.70, 0.78]) and 0.69 (95% CI: [0.66, 0.73]) (Figure 3; Table S5). Performance variability across demographics raises fairness concerns (Table S4).

Heterogeneity of the treatment effect

Investigating dose-response profiles conditioned on baseline physiologies reveals significant heterogeneity across patient subgroups. Figure 5 illustrates selected profiles for various clinical baselines, treatments, and outcomes. For instance, the causal model suggests patients with profile s94, featuring lower non-cardiorenal SOFA components, may achieve 4% improvement in hospital survival at fluid doses around 2 ml/kg/hour. Conversely, profiles s93 and s5, which exhibit higher mean blood pressures, show expected improvements below 2%. Similarly, we observe distinct responsiveness to medium vasopressor doses between 0.05 and 0.20 mcg/kg/min across subgroups, with clinical improvement rates varying from 0.5 to 17%, reflecting differences in acute severity such as urine output. Crucially, across these profiles, treatment benefits diminish or become harmful at higher doses. In certain scenarios, the model identifies effect plateaus, reflecting data limitations that prevent resolving fluid doses finer than 1.5 ml/kg/hour. Here, clinicians can leverage contextual information to determine optimal dosing. See Table S9 in Supplementary C for a complete clinical characterization of these examples.

Robustness to design choices

To rigorously quantify the robustness of our conclusions, we conducted an extensive sensitivity analysis encompassing 36 alternative scenarios by varying study designs, inclusion criteria, and causal graphs. In 33 scenarios, diverging from the model suggested doses was consistently associated with a reduced likelihood of survival and clinical improvement. The three exceptions involved a substantially reduced training set size, a simplified linear model, and the exclusion of patients receiving vasopressor therapy. Furthermore,

the difference in predictive performance between the causal model and the predictive baseline remained negligible across 76% of the examined scenarios. Additional details are provided in Tables S8, S7 and S6 in Supplementary B.

Discussion

Our proposed framework, evaluated on 3,136 ICU admissions across the MIMIC and eICU databases, provides a foundational example of how to introduce this structured causal methodology in intensive care. In such complex environments, causal models can inform decision-making through predictions robustly grounded in clinical, physiological, and operational knowledge. The causal graph ensures alignment between modeling assumptions and clinical reasoning, while enabling the estimation of hypothetical treatment effects directly from observational data. A notable finding was the complexity of the optimal policy space, defined as the combination of fluid intake and vasopressors doses, which revealed non-concave distributions rather than the simple U-shaped patterns that would unambiguously point to a single optimal policy. Accordingly, our causal model identifies a set of feasible policies where the clinician retains the final decision-making authority, a property we argue is not only inevitable given the complexity of the problem at hand, but also desirable in clinical practice.

Notably, by explicitly integrating causal reasoning into the modelling framework, the causal model uses approximately 65% fewer variables than conventional predictive approaches. This substantially reduces data requirements and computational complexity, potentially facilitating implementation, scalability, and deployment across diverse clinical environments while preserving clinically relevant decision support. Together with its superior

predictive performance and robustness under changing conditions, these findings highlight the potential of causal machine learning to support more reliable and generalizable clinical decision-making. This results in leaner, more reliable and auditable models with significantly lower energy consumption requirements, making causal models easier to deploy across healthcare facilities with varying degrees of AI capacity and data readiness, with tangible benefits for low- and middle-income countries and rural areas. These desirable properties are rooted in the process of constructing the causal graph itself. Constructing such a graph necessitates intentional, cross-functional collaboration between clinicians and data scientists, fostering the creation of joint teams with the shared competence required for the success of any AI application. Furthermore, although the resulting causal graph is necessarily a simplification and is potentially subject to error, its explicit nature allows it to be scrutinized and improved by other researchers. This transparency stands in contrast to black-box models, providing a shared representational language that facilitates incremental refinement and trust.

Our study has several limitations. First, the causal graph isolates haemodynamic physiology, neglecting interference from concurrent interventions like mechanical ventilation, sedation, or secondary diagnoses. Second, we assume data are missing at random, and variable aggregation may capture hospital operational biases rather than patient physiology, especially in the eICU dataset. Third, a significant predictive performance drop among Hispanic populations indicates fairness concerns and potential unmodeled mechanisms, addressable by incorporating sociodemographic factors or overlooked

mediators. Fourth, interventional validation findings remain susceptible to confounding by indication, survivor bias, immortal time bias, and time zero bias despite robust sensitivity analyses. Finally, the chosen statistical estimator may conservatively underestimate treatment effect sizes.

Constructing a causal graph through multidisciplinary consensus ensures that recommendations are grounded in physiological principles rather than incidental statistical associations. This structural approach enables the *Causal Artificial Intelligence Clinician* to simulate interventional scenarios, providing transparent and clinically interpretable insights for early haemodynamic management. By explicitly encoding domain knowledge, the framework establishes a robust methodology for estimating treatment effects and facilitating evidence-based decision-making in intensive care settings.

Contributors

GA: Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Visualization, Writing-original draft.
LA: Conceptualization, Investigation, Methodology, Supervision, Visualization, Writing-review & editing.
MC: Conceptualization, Investigation, Methodology, Supervision, Visualization, Writing-review & editing.
MZ: Conceptualization, Investigation, Methodology, Supervision, Visualization, Writing-review & editing.
MC and MZ jointly supervised this work.

Data Sharing Statement

Data used in this research was collected from publicly available databases, the MIMIC-IV and eICU databases. The procedure to access the MIMIC-IV is available at: `https://physionet.org/content/mimiciv/3.1/`. The procedure to access the eICU is available at: `https://physionet.org/content/eicu-crd/2.0/`. The MIMIC-IV dataset has been de-identified, and the institutional review boards of the Massachusetts Institute of Technology (0403000206) and Beth Israel Deaconess Medical Center (2001-P-001699/14) both approved the use of the database for research. Use of the eICU database is exempt from institutional review board approval due to the retrospective design, lack of direct patient intervention, and security schema, for which the re-identification risk was certified as meeting safe harbor standards by an independent privacy expert (Privacert, Cambridge, MA, Health Insurance Portability and Accountability Act Certification No. 1031219-2). Our work is fully reproducible, instructions to build the datasets and re-run our analysis is available at: `https://github.com/IDSIA/causal-ai-clinician`.

Declaration of interests

The authors declare no conflicts of interest related to this manuscript.

Acknowledgments

Figures and tables

		Overall	MIMIC	eICU
n		3136	1702	1434
Gender, n (%)	Female	1340 (42.7)	713 (41.9)	627 (43.7)
CCI , median [Q1,Q3]		4.0 [2.0,7.0]	5.0 [3.0,8.0]	4.0 [2.0,6.0]
SOFA , median [Q1,Q3]		5.0 [3.0,8.0]	5.0 [3.0,8.0]	6.0 [4.0,8.0]
Mechanical Ventilation, n (%)		1693 (54.0)	1025 (60.2)	668 (46.6)
Backdoors (First 6 Hours)				
Admission Age [Year(s)], median [Q1,Q3]		65.0 [53.0,75.0]	66.0 [53.0,77.0]	64.0 [53.0,74.0]
Diastolic Blood Pressure (Min) [mmHg], median [Q1,Q3]		45.0 [38.0,51.0]	46.0 [40.0,53.0]	43.0 [36.0,49.0]
Mean Blood Pressure (Min) [mmHg], median [Q1,Q3]		58.7 [52.0,63.7]	61.3 [56.0,66.0]	55.0 [48.0,61.0]
Heart Rate (Avg) [BPM], median [Q1,Q3]		93.1 [79.8,107.8]	92.0 [78.3,106.5]	94.5 [81.5,109.3]
Lactate (Max) [mmol/L], median [Q1,Q3]		3.6 [2.6,5.6]	3.8 [2.8,5.9]	3.4 [2.5,5.1]
Urine Output (Total) [ml/kg/hr], median [Q1,Q3]		0.7 [0.3,1.3]	0.7 [0.3,1.4]	0.6 [0.3,1.3]
Creatinine (Max) [mg/dL], median [Q1,Q3]		1.5 [1.0,2.4]	1.4 [1.0,2.3]	1.5 [1.0,2.4]
Clinical Descriptors (First 6 Hours)				
Residual SOFA , median [Q1,Q3]		1.7 [0.2,3.0]	1.0 [0.0,2.5]	2.0 [0.8,4.0]
WBC (Avg) [K/uL], median [Q1,Q3]		14.3 [9.4,19.6]	14.2 [9.5,19.5]	14.4 [9.2,19.9]
Temperature (Avg) [C], median [Q1,Q3]		36.8 [36.5,37.2]	36.8 [36.4,37.1]	36.8 [36.5,37.4]
Interventions (First 6 Hours)				
Norepinephrine Equivalent Dose (Total) [ug/kg/min], median [Q1,Q3]		0.1 [0.1,0.2]	0.1 [0.1,0.2]	0.1 [0.1,0.2]
Intravenous Fluids Dose (Avg) [ml/kg/hr], median [Q1,Q3]		1.5 [0.6,3.2]	1.4 [0.5,2.8]	1.7 [0.7,3.8]
Outcomes				
Clinical Improvement, n (%)		382 (12.2)	135 (7.9)	247 (17.2)
In-Hospital Survival, n (%)		2535 (80.8)	1266 (74.4)	1269 (88.5)

Table 1: Cohort Description. CCI: Charlson Comorbidity Index, SOFA Score: Sequential Organ Failure Assessment Score, Residual SOFA Score: SOFA score excluding renal and cardiovascular sections,GCS: Glasgow Coma Scale.

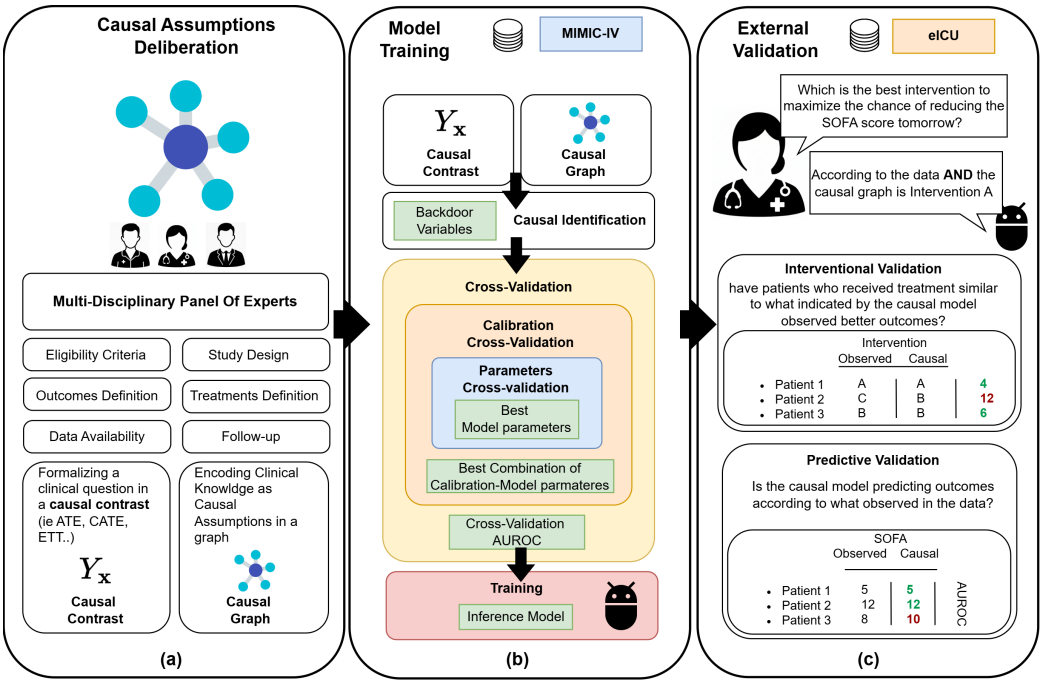


Figure 1: Visual summary of the core parts of the causal analysis.

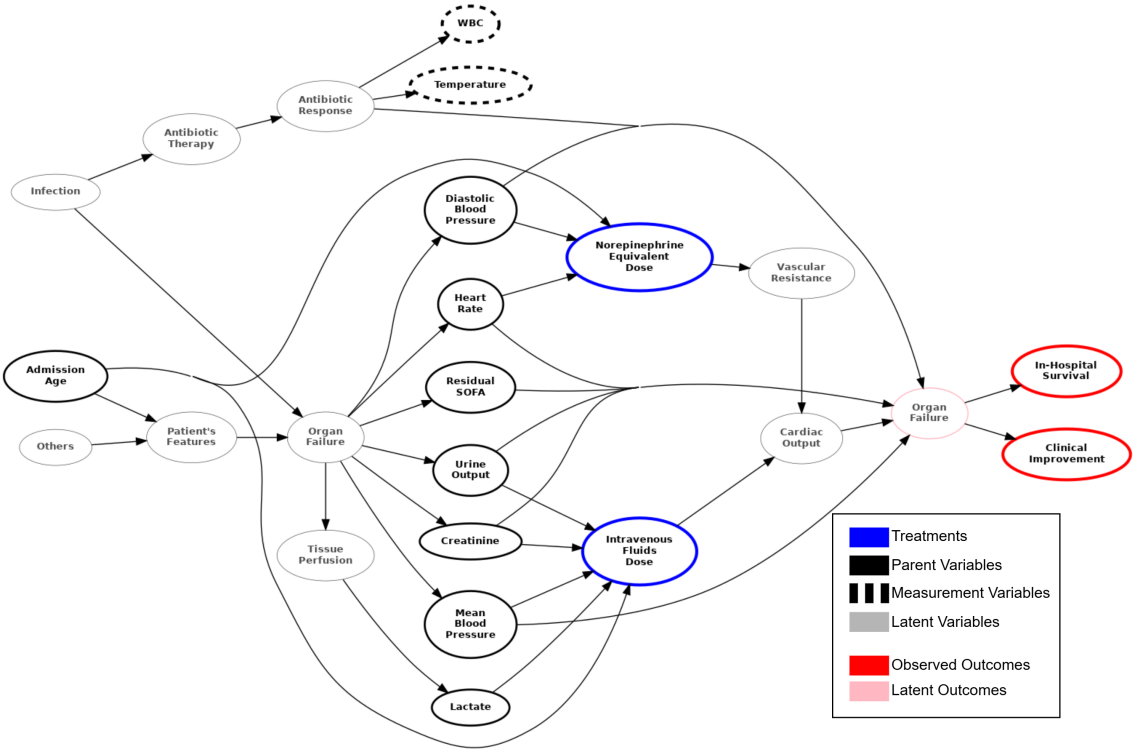


Figure 2: Causal Graph: DAG encoding the assumptions of our study. Outcomes are shown in red, decision variables in blue, backdoor variables in black, noisy proxies in dashed black, unmeasured variables in grey and unmeasured outcomes in pink.

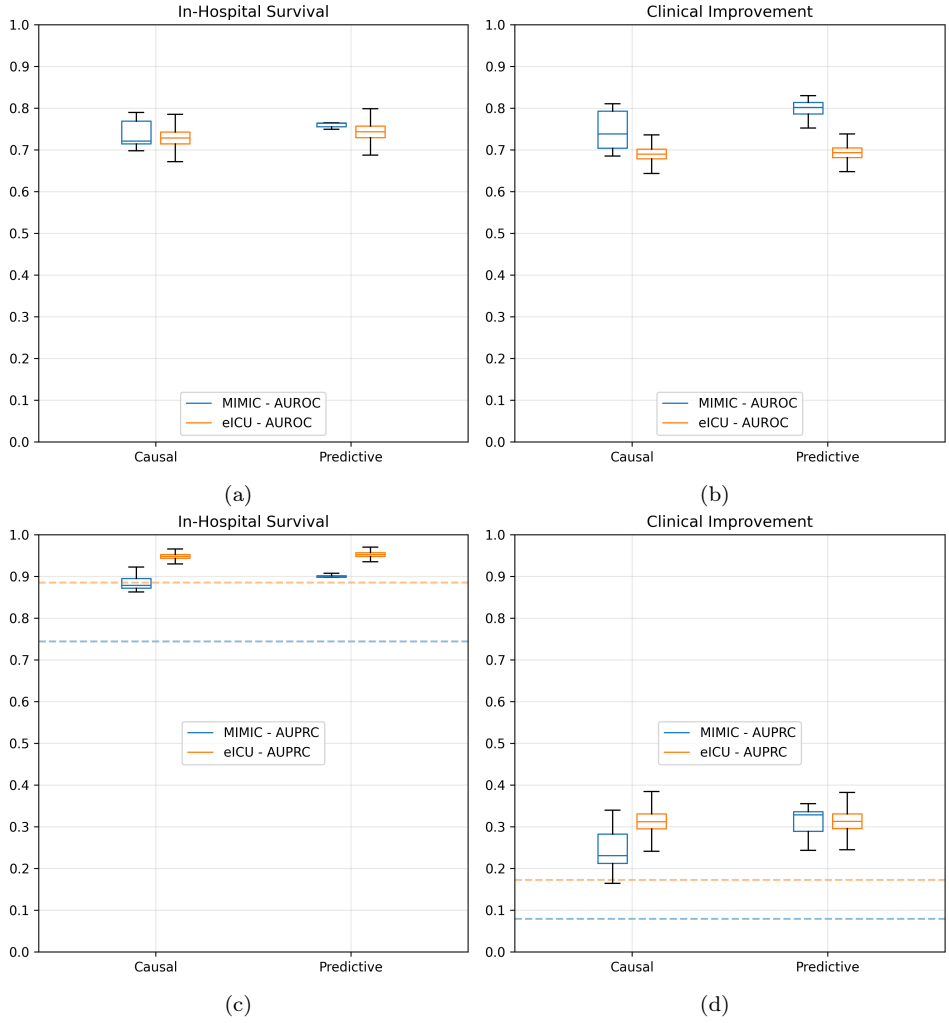
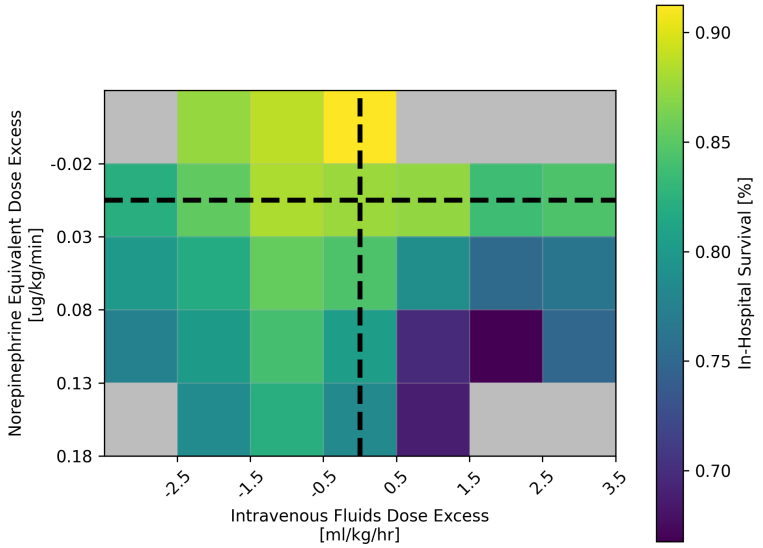
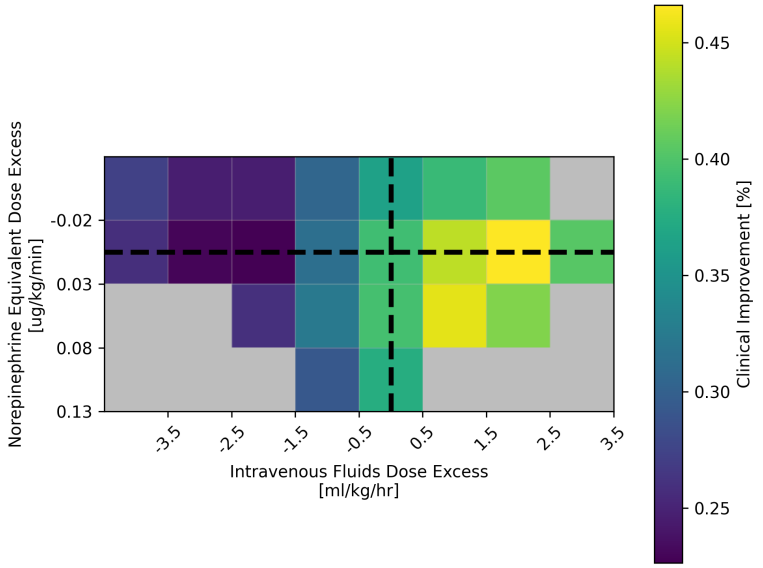


Figure 3: Causal and predictive model prediction performance the two study outcomes: in-hospital survival (a, c) and clinical improvement (b, d). Boxplots for AUROC (a, b) and AUPRC (c, d) evaluated in cross-validation on MIMIC (Blue) and in external validation on eICU (Orange). Dashed lines represent baseline prevalence.



(a)



(b)

Figure 4: External interventional validation: heatmaps representing outcomes prevalence in relation to the difference between observed prescribed dose in eICU and predicted optimal dose according to the causal model for (a) in-hospital survival and (b) clinical improvement. Lighter shades indicate groups of patients with a higher prevalence of better outcomes, while darker shades indicate poorer outcomes. The horizontal black dotted line represents optimal norepinephrine equivalent dose, the vertical dotted line represents optimal intravenous fluid dose. Their intersection represents the optimal combination of intervenable variables.

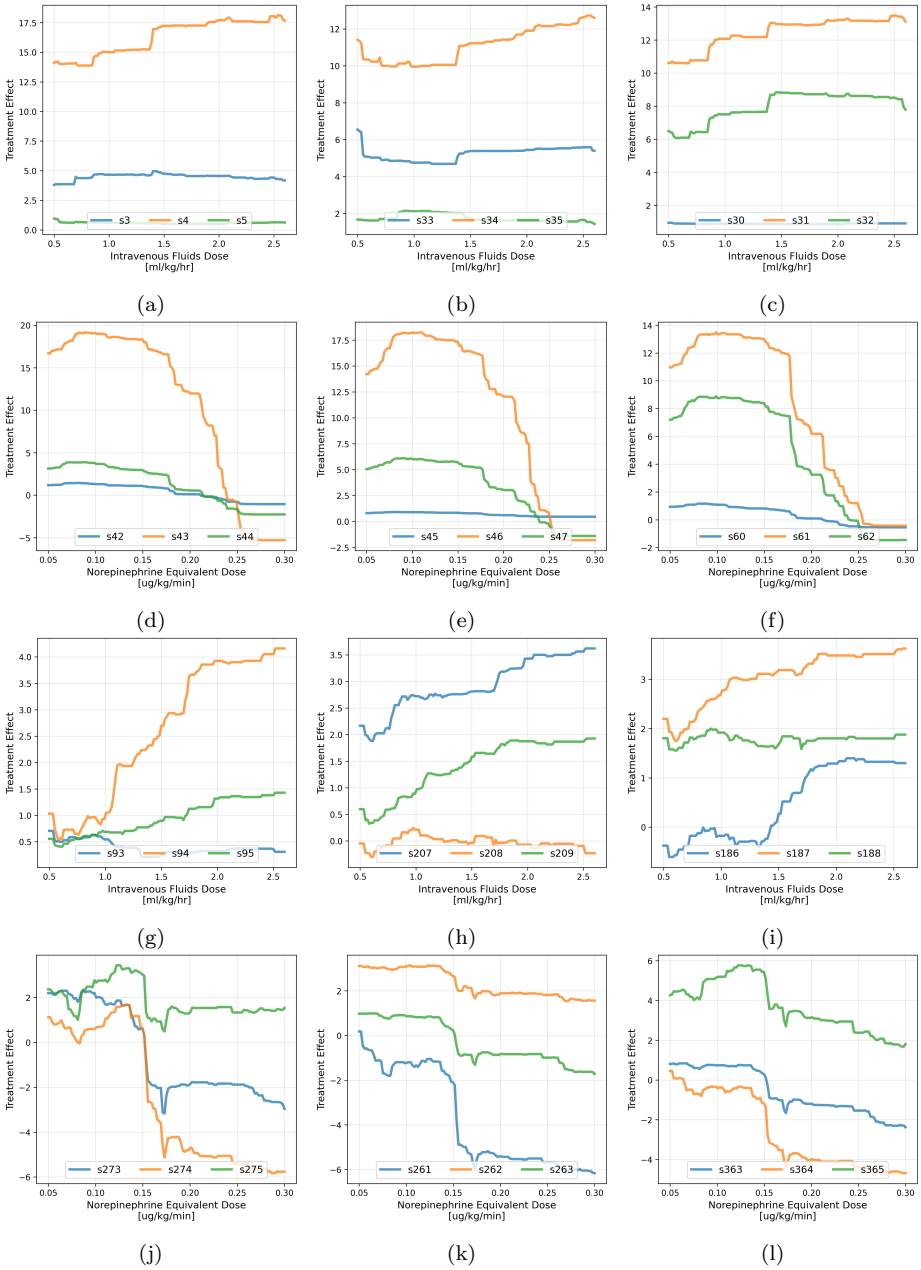


Figure 5: Simulated interventions: change in outcome predicted by the causal model (y-axis) as a function of simulated dose (x-axis). Positive values indicate a shift toward better prognosis, negative values a harmful treatment effect. Panels (a-c) and (d-f) show the treatment-response curves for intravenous fluid and norepinephrine doses with respect to clinical improvement, respectively; panels (g-i) and (j-l) for in-hospital survival. For each curve, the other treatment is considered fixed at the model-recommended optimal dose. Each panel shows three cohorts of patients with by different clinical profiles.

References

[1] Emanuel Rivers, Bryant Nguyen, Stacie Havstad, Julie Ressler, Alexandria Muzzin, Bernhard Knoblich, Edward Peterson, and Michael Tomlanovich. "Early Goal-Directed Therapy in the Treatment of Severe Sepsis and Septic Shock". In: *New England Journal of Medicine* 345.19 (2001), pp. 1368–1377. DOI: https://doi.org/10.1056/NEJMoa010307.

[2] Mitchell P Fink. "Cytopathic hypoxia: mitochondrial dysfunction as mechanism contributing to organ dysfunction in sepsis". In: *Critical care clinics* 17.1 (2001), pp. 219–237. DOI: https://doi.org/10.1016/s0749-0704(05)70161-5.

[3] Anand Kumar, Daniel Roberts, Kenneth E Wood, Bruce Light, Joseph E Parrillo, Satendra Sharma, Robert Suppes, Daniel Feinstein, Sergio Zanotti, Leo Taiberg, et al. "Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock". In: *Critical care medicine* 34.6 (2006), pp. 1589–1596. DOI: https://doi.org/10.1097/01.CCM.0000217961.75225.E9.

[4] Laura Evans, Andrew Rhodes, Waleed Alhazzani, Massimo Antonelli, Craig M Coopersmith, Craig French, Flavia R Machado, Lauralyn Mcintyre, Marlies Ostermann, Hallie C Prescott, et al. "Surviving Sepsis Campaign: International Guidelines for Management of Sepsis and Septic Shock 2021". In: *Critical Care Medicine* 49.11 (2021), e1063–e1143. DOI: https://doi.org/10.1097/CCM.0000000000005337.

[5] Mitchell M Levy, Laura E Evans, and Andrew Rhodes. "The surviving sepsis campaign bundle: 2018 update". In: *Intensive care medicine* 44.6 (2018), pp. 925–928. DOI: https://doi.org/10.1007/s00134-018-5085-0.

[6] Chairat Permpikul, Surat Tongyoo, Tanuwong Viarasilpa, Thavinee Trainarongsakul, Tipa Chakorn, and Suthipol Udompanturak. "Early use of norepinephrine in septic shock resuscitation (CENSER). A randomized trial". In: *American journal of respiratory and critical care medicine* 199.9 (2019), pp. 1097–1105. DOI: https://doi.org/10.1164/rccm.201806-1034OC.

[7] Christopher W Seymour, Jason N Kennedy, Shu Wang, Chung-Chou H Chang, Corrine F Elliott, Zhongying Xu, Scott Berry, et al. "Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis". In: *JAMA* 321.20 (2019), pp. 2003–2017. DOI: https://doi.org/10.1001/jama.2019.5791.

[8] Aleksandra Leligdowicz and Michael A Matthay. "Heterogeneity in sepsis: new biological evidence with clinical applications". In: *Critical care* 23.1 (2019), p. 80. DOI: https://doi.org/10.1186/s13054-019-2372-2.

[9] Reanimation ANDROMEDA-SHOCK-2 Investigators for the ANDROMEDA Research Network Spanish Society of Anesthesiology, Pain Therapy (SEDAR), Latin American Intensive Care Network (LIVEN), Glenn Hernandez, Gustavo A Ospina-Tascón, Eduardo Kattan, Miguel Ibarra-Estrada, Fernando Ramasco, Nicolás Orozco, Karla Ramos, Jose Luis Aldana, Giorgio Ferri, et al. "personalized hemodynamic resuscitation targeting capillary refill time in early septic shock: the ANDROMEDA-SHOCK-2 randomized clinical trial". In: *JAMA* 334.22 (2025), pp. 1988–1999.

[10] Victor B Talisa, Sachin Yende, Christopher W Seymour, and Derek C Angus. "Arguing for adaptive clinical trials in sepsis". In: *Frontiers in immunology* 9 (2018), p. 1502. DOI: https://doi.org/10.3389/fimmu.2018.01502.

[11] Natalja L Stanski and Hector R Wong. "Prognostic and predictive enrichment in sepsis". In: *Nature Reviews Nephrology* 16.1 (2020), pp. 20–31. DOI: https://doi.org/10.1038/s41581-019-0199-3.

[12] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. "MIMIC-IV, a freely accessible electronic health record dataset". In: *Scientific data* 10.1 (2023), p. 1. DOI: https://doi.org/10.1038/s41597-022-01899-x.

[13] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. "The eICU Collaborative Research Database, a freely available multi-center database for critical care research". In: *Scientific data* 5.1 (2018), pp. 1–13. DOI: https://doi.org/10.1038/sdata.2018.178.

[14] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. "The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care". In: *Nature medicine* 24.11 (2018), pp. 1716–1720. DOI: https://doi.org/10.1038/s41591-018-0213-5.

[15] Alexandre Kalimouttou, Jason N Kennedy, Jean Feng, Harvineet Singh, Suchi Saria, Derek C Angus, Christopher W Seymour, and Romain Pirracchio. "Optimal vasopressin initiation in septic shock: the OVISS reinforcement learning study". In: *Jama* 333.19 (2025), pp. 1688–1698. DOI: https://doi.org/doi:10.1001/jama.2025.3046.

[16] Hana Lee, Sky Qiu, Spencer Haupert, Gabriel K Innes, Tristan Naumann, Demissie Alemayehu, and Mark van der Laan. "Methodological and regulatory considerations for causal AI in drug development". In: *npj Digital Medicine* (2026). DOI: https://doi.org/10.1038/s41746-026-02477-w.

[17] Girish N Nadkarni, Jolion McGreevy, Nicholas Gavin, Robert Freeman, Bruce Darrow, Lisa Stump, David Reich, Isaac Kohane, and Ankit Sakhuja. "From prediction to navigation for artificial intelligence in medicine". In: *The Lancet* (2026). DOI: https://doi.org/10.1016/S0140-6736(26)00756-7.

[18] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[19] Marco Zaffalon, Alessandro Antonucci, Rafael Cabañas, David Huber, and Dario Azzimonti. "Efficient computation of counterfactual bounds". In: *International Journal of Approximate Reasoning* 171 (2024), p. 109111. DOI: https://doi.org/10.1016/j.ijar.2023.109111.

[20] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. *Double/debiased machine learning for treatment and structural parameters*. 2018. DOI: https://doi.org/10.1111/ectj.12097.

[21] Sema K Sgaier, Vincent Huang, and Grace Charles. "The case for causal AI". In: *Stanford Social Innovation Review* 18.3 (2020), pp. 50–55.

[22] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. "On Pearl's Hierarchy and the Foundations of Causal Inference". In: *Probabilistic and Causal Inference: The Works of Judea Pearl*. 1st ed. New York, NY, USA: Association for Computing Machinery, 2022, pp. 507–556. ISBN: 9781450395861. DOI: https://doi.org/10.1145/3501714.3501743. URL: https://doi.org/10.1145/3501714.3501743.

[23] Jin Tian and Judea Pearl. "A general identification condition for causal effects". In: *Aaai/iaai*. 2002, pp. 567–573. DOI: https://doi.org/10.5555/777092.777180.

[24] Md Osman Gani, Shravan Kethireddy, Riddhiman Adib, Uzma Hasan, Paul Griffin, and Mohammad Adibuzzaman. "Structural causal model with expert augmented knowledge to estimate the effect of oxygen therapy on mortality in the ICU". In: *Artificial Intelligence in Medicine* 137 (2023), p. 102493. DOI: https://doi.org/10.1016/j.artmed.2023.102493.

[25] Drago Plecko, Paul Secombe, Andrea Clarke, Amelia Fiske, Samarra Toby, Donisha Duff, David Pilcher, Leo Anthony Celi, Rinaldo Bellomo, and Elias Bareinboim. "An algorithmic approach for causal health equity: a look at race differentials in Intensive care unit (ICU) outcomes". In: *arXiv preprint arXiv:2501.05197* (2025). DOI: https://doi.org/10.48550/arXiv.2501.05197.

[26] J. Vogel, J. Cordier, and M. Filipovic. *Causal Effects and Optimal Policy Learning for Intensive Care Unit Discharge Decisions to Solve Hospital Process Bottlenecks*. Tech. rep. SSRN Working Paper Series, 2025.

[27] Michael Oberst and David Sontag. "Counterfactual Off-Policy Evaluation with Gumbel-Max Structural Causal Models". In: *International Conference on Machine Learning (ICML)*. 2019, pp. 4881–4890.

[28] Leland Gerson Neuberg. "Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000". In: *Econometric Theory* 19.4 (2003), pp. 675–685. DOI: https://doi.org/10.1017/S0266466603004109.

[29] Junzhe Zhang, Jin Tian, and Elias Bareinboim. "Partial counterfactual identification from observational and experimental data". In: *International conference on machine learning*. PMLR. 2022, pp. 26548–26558.

[30] Manabu Kuroki and Judea Pearl. "Measurement bias and effect restoration in causal inference". In: *Biometrika* 101.2 (2014), pp. 423–437. DOI: https://doi.org/10.1093/biomet/ast066.