The Causal Artificial Intelligence Clinician for early hemodynamic 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*

A. List of Abbreviations

- **AI**: Artificial Intelligence
- **ATE**: Average Treatment Effect
- **AUPRC**: Area Under the Precision-Recall Curve
- **AUROC**: Area Under the Receiver Operating Characteristic curve
- **CATE**: Conditional Average Treatment Effect
- **CI**: Confidence Interval
- **CCI**: Charlson Comorbidity Index
- **DAG**: Directed Acyclic Graph

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

- **DNR**: Do Not Resuscitate
- **ED**: Emergency Department
- **eICU**: eICU Collaborative Research Database
- **EHR**: Electronic Health Records
- **ER**: Emergency Room
- **ETT**: Effect of the Treatment on the Treated
- **GCS**: Glasgow Coma Scale
- **ICU**: Intensive Care Unit
- **IQR**: Interquartile Range
- **LMIC**: Low- and Middle-Income Countries
- **MAR**: Missing At Random
- **MIMIC**: Medical Information Mart for Intensive Care
- **ML**: Machine Learning
- **RCT**: Randomized Controlled Trial
- **SCM**: Structural Causal Model
- **SOFA**: Sequential Organ Failure Assessment
- **STEM**: Science, Technology, Engineering, and Mathematics
- **WBC**: White Blood Cell

B. Cohort details and commentary

Following the application of inclusion and exclusion criteria (Fig. S2), we obtained a final cohort comprising 3,136 intensive care unit admissions, split in 1,702 from the MIMIC and 1,434 from the eICU. Overall, our cohort exhibited a male majority (57.3 percent) and a high rate of invasive mechanical ventilation (54.0 percent). Baseline indicators demonstrated substantial comorbidity, with a median Charlson Comorbidity Index (CCI) of 4 (interquartile range [IQR] 2-7), and elevated admission severity, marked by a median SOFA score of 5 (IQR 3-8). Outcome assessment revealed an overall in-hospital survival rate of 80.0 percent with 12.2 percent of patients achieving clinical improvement (2-points reduction in SOFA score). Significant clinical heterogeneity was present between the training and external validation dataset. Patients sourced from the MIMIC database generally exhibited more severe conditions (odds ratio [OR] 1.40, 95% CI: [1.39, 1.41]) despite a lower unadjusted median SOFA score (MIMIC median 5, IQR 3-8 vs. eICU median 6, IQR 4-8). Furthermore, the MIMIC sub-cohort presented with higher comorbidity (CCI, OR 1.17, 95% CI: [1.16, 1.17]), higher recorded lactate levels (OR 1.06, 95% CI: [1.05, 1.07]), and higher vasopressor requirements (OR 1.22, 95% CI: [1.21, 1.23]). Consequently, MIMIC patients had a higher likelihood of poor outcomes, including lower in-hospital survival (74.4 vs. 88.8 percent; OR 0.76, 95% CI 0.75-0.77) and lower clinical improvement rates (7.9 vs. 17.6 percent; OR 0.87, 95% CI: [0.86, 0.88]). Conversely, patients in the eICU sub-cohort recorded higher non-cardiac and non-renal SOFA components (residual SOFA, OR 0.70, 95% CI: [0.69, 0.71]) and higher volumes of intravenous fluid intake (OR 0.93, 95% CI: [0.92, 0.94]). Comprehensive

descriptors are detailed in Table S1 and Table S2.

		Overall	MIMIC	eICU
n		3136	1702	1434
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]
Weight [Kgs], median [Q1,Q3]		79.3 [66.3,95.5]	78.4 [66.2,93.8]	80.4 [66.5,98.0]
CCI , median [Q1,Q3]		4.0 [2.0,7.0]	5.0 [3.0,8.0]	4.0 [2.0,6.0]
Gender, n (%)	Female	1340 (42.7)	713 (41.9)	627 (43.7)
Hypertension, n (%)		1256 (40.1)	557 (32.7)	699 (48.7)
Infection Location				
Abdominal Infection, n (%)		126 (4.0)	97 (5.7)	29 (2.0)
Respiratory Infection, n (%)		844 (26.9)	478 (28.1)	366 (25.5)
CNS Infection, n (%)		20 (0.6)	14 (0.8)	6 (0.4)
Urinary Infection, n (%)		409 (13.0)	201 (11.8)	208 (14.5)
Scores (First 6 Hours)				
SOFA , median [Q1,Q3]		5.0 [3.0,8.0]	5.0 [3.0,8.0]	6.0 [4.0,8.0]
Residual SOFA , median [Q1,Q3]		1.7 [0.2,3.0]	1.0 [0.0,2.5]	2.0 [0.8,4.0]
Vitals (First 6 Hours)				
GCS (Min), median [Q1,Q3]		15.0 [14.0,15.0]	15.0 [15.0,15.0]	14.0 [9.5,15.0]
Respiratory Rate (Avg) [BrPM], median [Q1,Q3]		20.5 [17.7,23.9]	20.6 [17.8,23.7]	20.5 [17.5,24.3]
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]
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]
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]
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]
Temperature (Avg) [C], median [Q1,Q3]		36.8 [36.5,37.2]	36.8 [36.4,37.1]	36.8 [36.5,37.4]
SpO2 (Min) [%], median [Q1,Q3]		94.0 [91.0,97.0]	95.0 [92.0,98.0]	93.0 [90.0,96.0]
Laboratory (First 6 Hours)				
Bilirubin (Max) [mg/dL], median [Q1,Q3]		0.7 [0.4,1.4]	0.7 [0.4,1.6]	0.7 [0.4,1.3]
Hemoglobin (Min) [g/dL], median [Q1,Q3]		10.6 [8.7,12.3]	10.5 [8.7,12.2]	10.7 [9.0,12.4]
Platelets (Min), median [Q1,Q3]		180.0 [120.0,255.0]	177.0 [118.0,256.0]	187.0 [125.0,253.0]
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]
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]
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]
Interventions				
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]
Mechanical Ventilation, n (%)		1693 (54.0)	1025 (60.2)	668 (46.6)
Others				
Admission Time, n (%)	00-06	754 (24.0)	335 (19.7)	419 (29.2)
	06-12	582 (18.6)	319 (18.7)	263 (18.3)
	12-18	781 (24.9)	505 (29.7)	276 (19.2)
	18-24	1019 (32.5)	543 (31.9)	476 (33.2)
Admission Year, n (%)	2011 - 2014	398 (12.7)	398 (23.4)	0 (0.0)
	2014 - 2017	398 (12.7)	398 (23.4)	0 (0.0)
	2017 - 2020	2101 (67.0)	667 (39.2)	1434 (100.0)
	2020 - 2023	595 (19.0)	595 (35.0)	0 (0.0)
Unit, n (%)	Cardiac	552 (17.6)	247 (14.5)	305 (21.3)
	Medical-Surgical	2245 (71.6)	1188 (69.8)	1057 (73.7)
	Neuro-Trauma	339 (10.8)	267 (15.7)	72 (5.0)
Post-vasotherapy MBP readings , n (%)		403 (12.9)	375 (22.0)	28 (2.0)
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 S1: Full set of features accessible to the predictive model plus outcomes. 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, WBC: White Blood Cell Count, SpO2: Blood oxygen saturation, MBP: Mean Blood Pressure. In bold variables were used by the causal model.

Feature	**Odds**
	Median [95% CI]
Admission Year:2017 - 2020	1.82 [1.76,1.89]
Admission Year:2011 - 2014	1.58 [1.54,1.63]
Mechanical Ventilation	1.57 [1.53,1.61]
GCS	1.47 [1.46,1.49]
SOFA	1.40 [1.39,1.41]
Clinical Stability	1.36 [1.34,1.38]
Post-Vasotherapy MBP	1.35 [1.33,1.37]
Unit:Neuro-Trauma	1.28 [1.26,1.30]
Norepinephrine Equivalent Dose	1.22 [1.21,1.23]
Hemoglobin	1.22 [1.20,1.22]
Mean Blood Pressure	1.18 [1.17,1.18]
Admission Time:12-18	1.17 [1.16,1.18]
CCI	1.17 [1.16,1.18]
Urine Output	1.16 [1.15,1.17]
Abdominal Infection	1.08 [1.07,1.08]
Lactate	1.06 [1.05,1.07]
Respiratory Infection	1.06 [1.05,1.06]
Gender	1.05 [1.05,1.05]
Admission Year: Other	1.05 [1.05,1.05]
SpO2	1.04 [1.04,1.04]
Admission Time:06-12	1.04 [1.03,1.04]
CNS Infection	1.01 [1.01,1.01]
Heart Rate	1.00 [1.00,1.00]
Platelets	1.00 [1.00,1.00]
WBC	1.00 [1.00,1.00]
Respiratory Rate	0.99 [0.99,1.00]
Weight	0.99 [0.99,0.99]
Admission Age	0.99 [0.99,0.99]
Bilirubin	0.98 [0.98,0.99]
Urinary Infection	0.97 [0.97,0.97]
Admission Time:18-24	0.97 [0.96,0.97]
Intravenous Fluids Dose	0.93 [0.92,0.93]
Unit:Medical-Surgical	0.90 [0.90,0.91]
Diastolic Blood Pressure	0.90 [0.89,0.90]
Clinical Improvement	0.87 [0.86,0.88]
Unit:Cardiac	0.86 [0.85,0.87]
Admission Time:00-06	0.84 [0.84,0.85]
Creatinine	0.77 [0.76,0.77]
In-Hospital Survival	0.76 [0.75,0.77]
Hypertension	0.75 [0.73,0.76]
Residual SOFA	0.70 [0.69,0.71]
Temperature	0.67 [0.66,0.68]
Admission Year:2014 - 2017	0.35 [0.33,0.36]

Table S2: Odds of a patient belonging to the MIMIC cohort as against eICU

Feature	**MIMIC (%)**	**eICU (%)**
Admission Age	0.0	0.0
Weight	0.0	0.0
CCI	0.0	0.0
Gender	0.0	0.0
Hypertension	0.0	0.0
SOFA	0.4	0.0
Residual SOFA	0.4	0.0
GCS	1.8	20.3
Respiratory Rate	0.1	0.9
Heart Rate	0.2	0.1
Mean Blood Pressure	0.2	0.0
Diastolic Blood Pressure	0.2	0.5
Urine Output	10.6	31.9
Temperature	5.6	1.9
SpO2	0.4	2.0
Abdominal Infection	0.0	0.0
Respiratory Infection	0.0	0.0
CNS Infection	0.0	0.0
Urinary Infection	0.0	0.0
Bilirubin	21.8	52.7
Hemoglobin	0.5	34.8
Platelets	0.5	37.7
WBC	0.5	37.7
Creatinine	0.4	30.8
Lactate	28.3	43.4
Norepinephrine Equivalent Dose	29.6	77.5
Intravenous Fluids Dose	0.0	0.0
Mechanical Ventilation	0.0	0.0
Admission Time	0.0	0.0
Admission Year	0.0	0.0
Unit	0.0	0.0
Post-Vasotherapy MBP	0.0	0.0
Clinical Stability	0.0	0.0
Clinical Improvement	0.0	0.0
In-Hospital Survival	0.0	0.0

Table S3: Missing rates across MIMIC and eICU variables.

B.1. Fairness Analysis Details

Category	Subgroup	Causal	Predictive
Gender	Female	-0.01 [-0.01,-0.01]	0.03 [0.03,0.04]
Race and Ethnicity	Asian		
	Black	0.02 [0.02,0.03]	0.06 [0.06,0.06]
	Hispanic	-0.16 [-0.17,-0.16]	-0.06 [-0.06,-0.05]
	Other	-0.07 [-0.07,-0.07]	-0.02 [-0.02,-0.01]

(a)

Category	Subgroup	Causal	Predictive
Gender	Female	-0.06 [-0.06,-0.06]	-0.05 [-0.05,-0.05]
Race and Ethnicity	Asian	-0.12 [-0.14,-0.11]	-0.11 [-0.13,-0.10]
	Black	-0.04 [-0.04,-0.03]	-0.01 [-0.01,-0.00]
	Hispanic	-0.08 [-0.09,-0.08]	0.03 [0.02,0.03]
	Other	0.01 [0.01,0.01]	0.03 [0.03,0.04]

(b)

Table S4: AUROC parity test. Gap in AUROC compared to Males for gender and Whites for race and ethnicity for (S4a) In-Hospital Survival and (S4b) Clinical Improvement. Values presented as Median [95% CI].

B.2. Cross-Validation and External Validation Details

	Clinical Improvement		In-Hospital Survival	
	MIMIC	eICU	MIMIC	eICU
Causal	0.74 [0.69,0.81]	0.69 [0.66,0.72]	0.72 [0.70,0.79]	0.73 [0.69,0.77]
Predictive	0.80 [0.76,0.83]	0.69 [0.66,0.73]	0.76 [0.75,0.76]	0.74 [0.70,0.78]

(a)

	Clinical Improvement		In-Hospital Survival	
	MIMIC	eICU	MIMIC	eICU
Causal	0.23 [0.17,0.33]	0.31 [0.26,0.37]	0.88 [0.86,0.92]	0.95 [0.93,0.96]
Predictive	0.33 [0.25,0.35]	0.31 [0.26,0.37]	0.90 [0.89,0.91]	0.95 [0.94,0.96]

(b)

Table S5: Predictive performance.(S5a) AUROC and (S5b) AUPRC in cross-validation (MIMIC) and external validation (eICU) for the studys' endpoints. Values presented as Median [95% CI].

B.3. Sensitivity Analysis Details and Commentary

For vasopressors, when including DNR patients, any additional unit deviation from the causal suggestion is associated with an 11.81 (95% CI: [11.52, 12.15]) increase in the likelihood of in-hospital death. Conversely, when reducing the cohort to only patients without missing lactates, this odd diminishes to 1.83. When focusing on the other endpoint, clinical improvement, odds for vasopressors are generally lower, with reported highest odds at 9.30 for a cohort with lower blood pressure inclusion threshold, at 60 mmHg instead of 65 mmHg, and lowest at 1.31, again for patients with no missing lactates. Intravenous fluid odds show less variance where, in all cases in which deviation from causal suggestions is associated with worsening of outcomes, values range between 1.01 and 1.15. When looking at survival, intravenous fluid deviations are lowest in the cohort where minimum ICU stay is lowered to 18 hours, at 1.01 (95% CI: [1.01, 1.01]), and highest on the cohort with

no missing lactates, at 1.06 (95% CI: [1.06, 1.06]). On clinical improvement, lowest odds,at 1.06, appears when using a logistic regression instead of gradient boosting for the CATE model and the highest, at 1.15, when using additional demographic backdoors.

Criteria		In-Hospital Survival		Causal	Predictive	Gap
		n	%	Median [95% CI]	Median [95% CI]	Median [95% CI]
Baseline	MIMIC	1702	.74	0.72 [0.70,0.79]	0.76 [0.75,0.76]	-0.03 [-0.07,0.03]
	eICU	1434	.88	0.73 [0.69,0.77]	0.74 [0.70,0.78]	-0.01 [-0.03,0.00]
Baseline, excluding No DNR	MIMIC	1789	.74	0.74 [0.68,0.77]	0.76 [0.73,0.78]	-0.03 [-0.06,0.01]
	eICU	2072	.74	0.72 [0.70,0.75]	0.75 [0.72,0.77]	-0.02 [-0.03,-0.01]
Baseline, excluding MAP $\leq$ 65 mmHg	MIMIC	1198	.71	0.71 [0.69,0.75]	0.74 [0.68,0.78]	-0.03 [-0.04,0.02]
	eICU	322	.81	0.72 [0.64,0.79]	0.77 [0.70,0.83]	-0.05 [-0.09,-0.02]
Baseline plus mention of sepsis in admission notes	MIMIC	264	.79	0.54 [0.47,0.62]	0.59 [0.53,0.76]	-0.06 [-0.21,0.04]
	eICU	946	.89	0.62 [0.56,0.68]	0.60 [0.54,0.66]	0.03 [-0.04,0.10]
Baseline, but MAP $\leq$ 60 mmHg instead of 65	MIMIC	1462	.73	0.74 [0.66,0.76]	0.75 [0.71,0.78]	-0.03 [-0.09,0.00]
	eICU	1163	.88	0.75 [0.70,0.79]	0.76 [0.72,0.80]	-0.01 [-0.04,0.01]
Baseline with raw lactate $\geq$ 2 mmol/L	MIMIC	1221	.71	0.76 [0.68,0.77]	0.75 [0.71,0.79]	-0.03 [-0.04,0.03]
	eICU	811	.87	0.73 [0.68,0.78]	0.75 [0.70,0.79]	-0.01 [-0.04,0.01]
Baseline plus auxilliary demographic variables	MIMIC	1702	.74	0.73 [0.70,0.77]	0.75 [0.73,0.79]	-0.03 [-0.06,0.04]
	eICU	1434	.88	0.73 [0.69,0.77]	0.74 [0.70,0.78]	-0.01 [-0.03,0.00]
Baseline but at least 24 hours of ICU stay	MIMIC	1791	.74	0.78 [0.69,0.79]	0.76 [0.73,0.82]	0.01 [-0.08,0.06]
	eICU	1563	.88	0.73 [0.71,0.74]	0.75 [0.74,0.76]	-0.02 [-0.04,-0.01]
Baseline, Logistic Regression as estimator	MIMIC	1702	.74	0.73 [0.66,0.74]	0.73 [0.69,0.78]	-0.04 [-0.07,0.04]
	eICU	1434	.88	0.70 [0.68,0.72]	0.72 [0.71,0.74]	-0.02 [-0.05,-0.01]
Baseline but at least 18 hours of ICU stay	MIMIC	1908	.74	0.77 [0.72,0.81]	0.79 [0.77,0.80]	-0.02 [-0.05,0.01]
	eICU	1674	.87	0.75 [0.71,0.78]	0.77 [0.73,0.80]	-0.02 [-0.04,-0.00]

Table S6: Sensitivity Analysis for the in-hospital survival endpoint: AUROC for different exclusion and inclusion criteria and choices of backdoors. The difference, in terms of AUROC, between the causal model and predictive model is presented under the gap column.

Criteria		n	Clinical Improvement %	Causal Median [95% CI]	Predictive Median [95% CI]	Gap Median [95% CI]
Baseline	MIMIC	1702	.08	0.74 [0.69,0.81]	0.80 [0.76,0.83]	-0.05 [-0.12,0.02]
	eICU	1434	.17	0.69 [0.66,0.72]	0.69 [0.66,0.73]	-0.00 [-0.03,0.02]
Baseline, excluding No DNR	MIMIC	1789	.08	0.75 [0.73,0.81]	0.78 [0.74,0.79]	0.02 [-0.06,0.04]
	eICU	2072	.16	0.69 [0.66,0.72]	0.68 [0.65,0.71]	0.01 [-0.02,0.03]
Baseline, excluding MAP $\leq$ 65 mmHg	MIMIC	1198	.1	0.74 [0.68,0.76]	0.78 [0.77,0.81]	-0.04 [-0.13,-0.02]
	eICU	322	.25	0.69 [0.63,0.76]	0.69 [0.62,0.75]	0.01 [-0.04,0.05]
Baseline plus mention of sepsis in admission notes	MIMIC	264	.09	0.66 [0.52,0.90]	0.63 [0.48,0.84]	0.06 [-0.28,0.43]
	eICU	946	.18	0.66 [0.61,0.70]	0.66 [0.61,0.70]	-0.00 [-0.02,0.01]
Baseline, but MAP $\leq$ 60 mmHg instead of 65	MIMIC	1462	.09	0.74 [0.72,0.84]	0.81 [0.71,0.86]	-0.08 [-0.14,0.08]
	eICU	1163	.17	0.69 [0.65,0.73]	0.70 [0.66,0.74]	-0.01 [-0.04,0.01]
Baseline with raw lactate $\geq$ 2 mmol/L	MIMIC	1221	.08	0.73 [0.70,0.76]	0.79 [0.68,0.83]	-0.07 [-0.10,0.08]
	eICU	811	.18	0.68 [0.64,0.72]	0.70 [0.65,0.75]	-0.02 [-0.05,0.01]
Baseline plus auxilliary demographic variables	MIMIC	1702	.08	0.75 [0.69,0.76]	0.79 [0.76,0.86]	-0.06 [-0.14,-0.02]
	eICU	1434	.17	0.69 [0.66,0.72]	0.69 [0.66,0.73]	-0.00 [-0.03,0.03]

Table S7: Sensitivity Analysis for the clinical improvement endpoint: AUROC for different exclusion and inclusion criteria and choices of backdoors. The difference, in terms of AUROC, between the causal model and predictive model is presented under the gap column. Please note that sensitivity analysis with shorter outcomes horizons (ie at 18 and 24 hours, see Table S6) were not computed since not enough time passed to re-calculate the SOFA score.

Criteria	**Therapy**	**Lack of In-Hospital Survival** Median [95% CI]	**Lack of Clinical Improvement** Median [95% CI]
Baseline	Norepinephrine Equivalent	5.61 [5.44,5.78]	6.33 [6.17,6.50]
	Intravenous Fluid Therapy	1.02 [1.02,1.02]	1.14 [1.14,1.14]
Baseline, excluding No DNR	Norepinephrine Equivalent	11.81 [11.52,12.15]	3.80 [3.70,3.91]
	Intravenous Fluid Therapy	1.02 [1.02,1.02]	1.10 [1.10,1.10]
Baseline, excluding MAP $\leq$ 65 mmHg	Norepinephrine Equivalent	1.93 [1.88,1.99]	5.62 [5.48,5.81]
	Intravenous Fluid Therapy	1.02 [1.02,1.03]	0.96 [0.96,0.97]
Baseline plus mention of sepsis in admission notes	Norepinephrine Equivalent	0.29 [0.28,0.30]	1.41 [1.38,1.45]
	Intravenous Fluid Therapy	1.04 [1.04,1.04]	1.04 [1.03,1.04]
Baseline, but MAP $\leq$ 60 mmHg instead of 65	Norepinephrine Equivalent	3.59 [3.49,3.70]	9.30 [9.06,9.56]
	Intravenous Fluid Therapy	1.04 [1.04,1.04]	1.08 [1.08,1.09]
Baseline with raw lactate $\geq$ 2 mmol/L	Norepinephrine Equivalent	1.83 [1.79,1.88]	1.31 [1.28,1.34]
	Intravenous Fluid Therapy	1.06 [1.06,1.06]	1.13 [1.13,1.13]
Baseline plus auxiliary demographic variables	Norepinephrine Equivalent	5.19 [5.01,5.36]	5.10 [4.94,5.26]
	Intravenous Fluid Therapy	1.02 [1.02,1.03]	1.15 [1.15,1.15]
Logistic Regression as estimator	Norepinephrine Equivalent	5.15 [4.99,5.32]	1.49 [1.45,1.53]
	Intravenous Fluid Therapy	0.97 [0.97,0.97]	1.06 [1.06,1.06]
At least 18 hours of ICU stay	Norepinephrine Equivalent	7.57 [7.34,7.81]	
	Intravenous Fluid Therapy	1.01 [1.01,1.01]	
At least 24 hours of ICU stay	Norepinephrine Equivalent	3.90 [3.78,4.02]	
	Intravenous Fluid Therapy	1.01 [1.01,1.01]	

B.4. Study Design and ML Pipeline

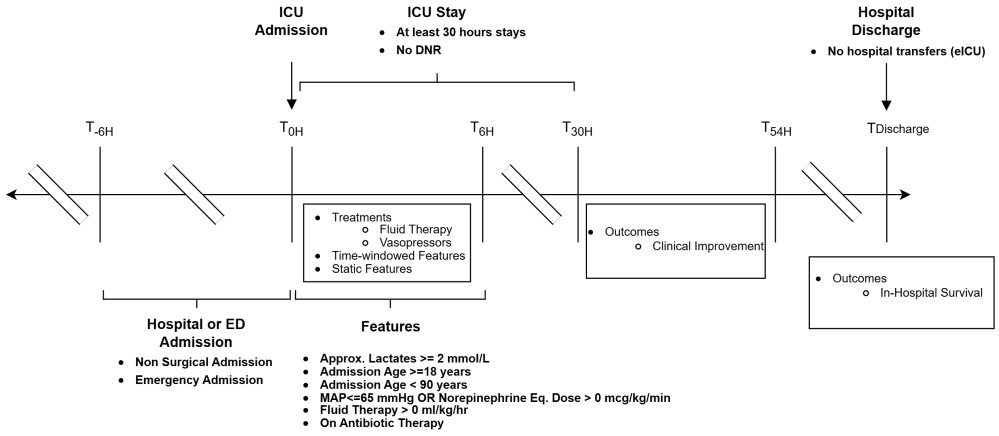


Figure S1: Study Design

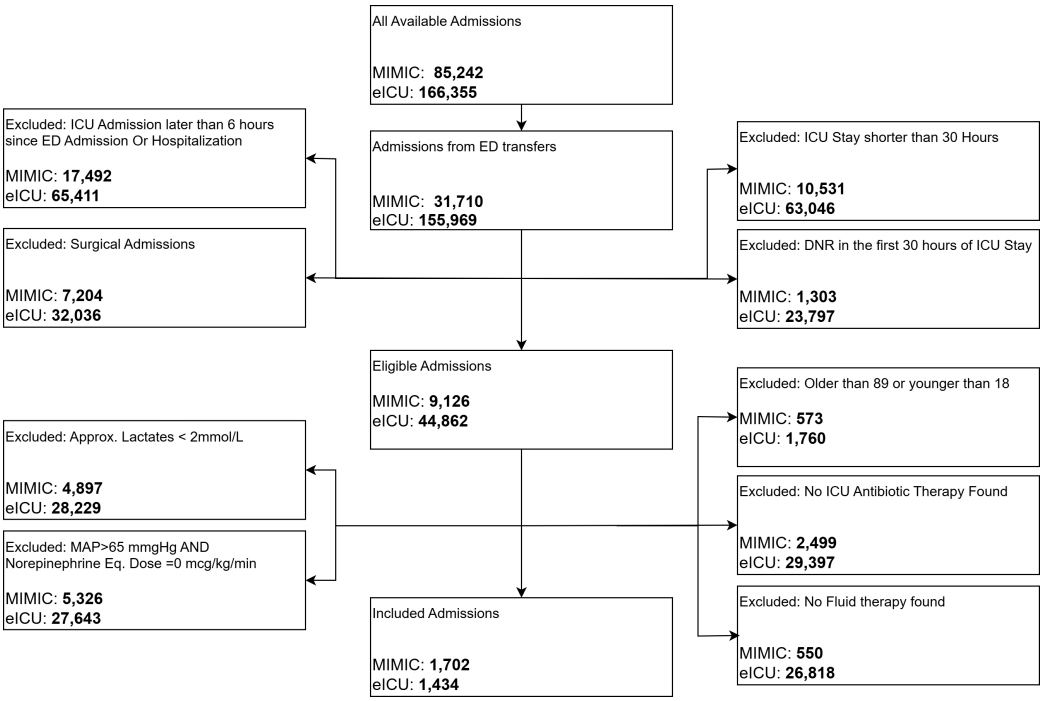


Figure S2: Inclusion and exclusion criteria.

C. Causal Inference Details

C.1. Methodological details

Causal inference enables ML models to move beyond correlation and differentiate causation from mere statistical association. This differentiation is guided by a set of assumptions derived from domain knowledge, which in our case draws on existing evidence about sepsis and ICU dynamics. While many established causal frameworks exist, each encoding the causal assumptions differently, our framework of choice are Structural Causal Models (SCMs). SCMs are causal graphical model where causal assumptions are represented as directed arrows in a directed acyclic graph (DAG). This graphical representation acts as a navigational map governed by the strong theoretical foundations of the do-calculus, which determines the possibility of robustly quantifying interventional "what-if" simulation scenarios. Beyond their theoretical grounding, the causal graph offers a natural, visually intuitive language that fosters multidisciplinarity by enabling healthcare professionals, STEM practitioners, and data scientists to reason using a shared, common representation. In our case, the resulting causal graph, is shown in Fig. 2, and the clinical "what-if" question we aim to quantify is:

"What is the probability of survival (Y) if we administer a dosage of norepinephrine (N) and intravenous fluid (F) in the early hours following ICU admission in septic shock patients sharing clinical baseline (X)?"

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 $\mathbf{X}$, equivalent in principle to performing an RCT. The CATE therefore represents the change in the probability of in-hospital survival or clinical improvement when delivering a combination of norepinephrine and fluids compared to no treatment. We employed the *Backdoor Criterion* to identify the features to include in the model. The Backdoor criterion can be verified from the causal graph alone and states that by adjusting for all parents $\mathbf{Z}$ of F and N, which correspond to a subset of $\mathbf{X}$, the causal query of interest can be approximated from the data[8](for a rigorous formulation of the Backdoor definition, please refer to corollary 3 in[2]). In our case, as shown in Fig. S3, the criterion is satisfied and the CATE can be estimated by using

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

where **W** are the remaining clinical descriptors in **X** that are not in **Z**.

The CATE can then be estimated in many different ways.[6,5,1,9] In our study, we adopt gradient boosting trees, an approach equivalent to an S-learner in meta-learner terminology[6]). A formal graphical representation of the causal assumptions among **Z**, **X**, **W**, N and F is presented in Fig. S3 in B.4.

C.2. Causal Graph Derivation Commentary

To derive the causal graph represented in Fig. 2, we focused primarily on hemodynamic stability and the role of intravenous fluids administration and vasopressors dosage in the early hours of ICU. While we acknowledge that this constitutes a simplification of reality, it provides a structural rationale grounded in the clinical decision process. The graph incorporates both observed variables, those we retrieved from both the MIMIC and eICU datasets, and unmeasured ones, which either we could not reliably extract or were not available in the Electronic Health Records (EHR) (shown in grey in Fig. 2).

The derivation of the causal graph was central to defining the blueprint of our causal model and represents a major contributing factor in the results we obtained. Consensus on the causal graph was built in collaboration with expert causal methodologists, data scientists, and physicians, iteratively verifying that the encoded assumptions were clinically sound, that variables could be estimated from the data, and that causal backdoors could be approximated from the current assumptions.

The graph presented in Fig. 2 can be broadly read from left to right in chronological order, starting from infections acquired before Emergency

Department (ED) admission to early antibiotic choices up until early treatments decisions in the ICU. Starting from the patient's ICU admission, we assume that appropriate antibiotics and fluid intake therapy were administered during the ER stay, as no data are available to reliably estimate them. From this points, we trace the clinician's decision process: the primary target is to stabilize tissue perfusion, which corresponds, as per guidelines, to setting a hemodynamic target of 65 mmHg of mean blood pressure. This can be achieved by acting on two levers: increasing cardiac output with fluids and increasing vascular resistance with vasopressors. Treatment doses are decided based on the progression of organ failure: norepinephrine equivalents are titrated based on signals of contractility, heart rate, and diastolic blood pressure. Renal function (via creatinine and urine output) and tissue perfusion (via lactates and mean blood pressure) inform clinicians on proper fluid dosing. In both cases, the patient's frailty is accounted for through their age as a proxy. The patient's response to infection is additionally captured through temperature and WBC count, which does not carry causal relevance in our model but enables stratification by approximate infection state.

The predictive model, used to benchmark the causal model's predictive performance, was built by including a broader set of routinely collected clinical and administrative markers from the ICU, in addition to all measured variables included in the causal graph. A full list of predictors is presented in Table S1.

C.3. Model Training Details

Training was performed via doubly nested cross-validation, where the inner loop iteratively sampled hyperparameters while the outer loop identified

optimal calibration parameters. Model calibration was essential to represent model predictions as probabilities, which was necessary for estimating treatment effects. In the final outer scoring cross-validation loop, AUROCs and AUPRCs were estimated on the training data, with AUROC selected as primary metric for its comparative properties.[7] In our experiment, for each outcome, we used 5 folds for the external scoring loop, 5 folds for the calibration loop, and 5 folds for the inner hyperparameter loop, each repeated for all 500 different hyperparameter combinations. AUROCs and AUPRCs in external validation were estimated over 2,000 bootstraps.

C.4. Optimal Intervention Estimation and Validation Details

The optimal treatment policy, defined as the combination of fluid and norepinephrine equivalent doses yielding the highest estimated treatment effect for a given patient, was estimated as

$$n^*, f^* = \underset{n,f}{\operatorname{argmax}} \quad \text{CATE}((n, f), \mathbf{X}),$$

where the optimization was performed across a discrete mesh grid of doses. The mesh grid's lower and upper limits were defined by the 5th and 95th percentiles observed in the training dataset. For each intervenable variable, i.e., norepinephrine and intravenous fluids, we examined 32 equidistant values in the mesh, equaling 1,024 different treatment combinations.

To validate the estimated policies, we compared patient outcomes in the external validation dataset according to the deviation between the administered treatment and the one suggested by the causal model. The rationale is that an informative policy should yield better outcomes for observed treatments aligning with model suggestions and worse prognoses for those

diverging from them. To quantify this, we estimated the odds of worse outcomes for every absolute unit increase in excess norepinephrine equivalent dose and fluid intake, computed as the average of 1,000 bootstraps over 10 repeated 10-fold cross-validations on the external validation set. As a visual aid, we computed the differences between the observed and optimal doses of both norepinephrine equivalent and fluid intake, for each patient in external validation. We then grouped patients by deviation magnitude and estimated the average prevalence of both outcomes within each group. Deviations were discretized in bins of 0.05 mcg/kg/min and 1 ml/kg/hr for norepinephrine and fluid intake, respectively. Outcome prevalence within each bin was estimated over 1,000 bootstraps. To account for aleatoric uncertainty arising from documentation and aggregation biases in the measured interventions, we introduced a normal noise term with a standard deviation equal to the bin size. This procedure increases variance but better reflects our confidence in the quality of the documentation process. Finally, to assess heterogeneity of the treatment effect across varying patient baselines, we estimated treatment effects conditioned on covariates defined across a grid of the input space. We then broadly clustered profiles by common treatment-response trends to identify common trends. The most clinically informative examples were selected to illustrate the heterogeneous effects of vasopressors and fluid intake on in-hospital survival and clinical improvement.

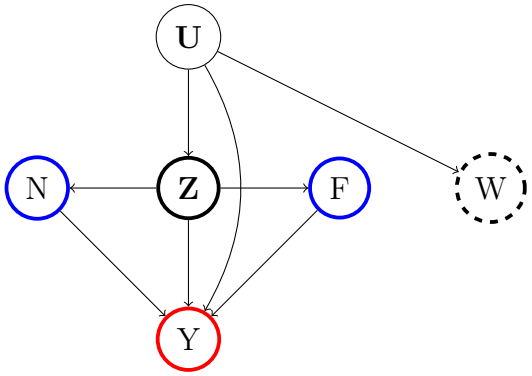
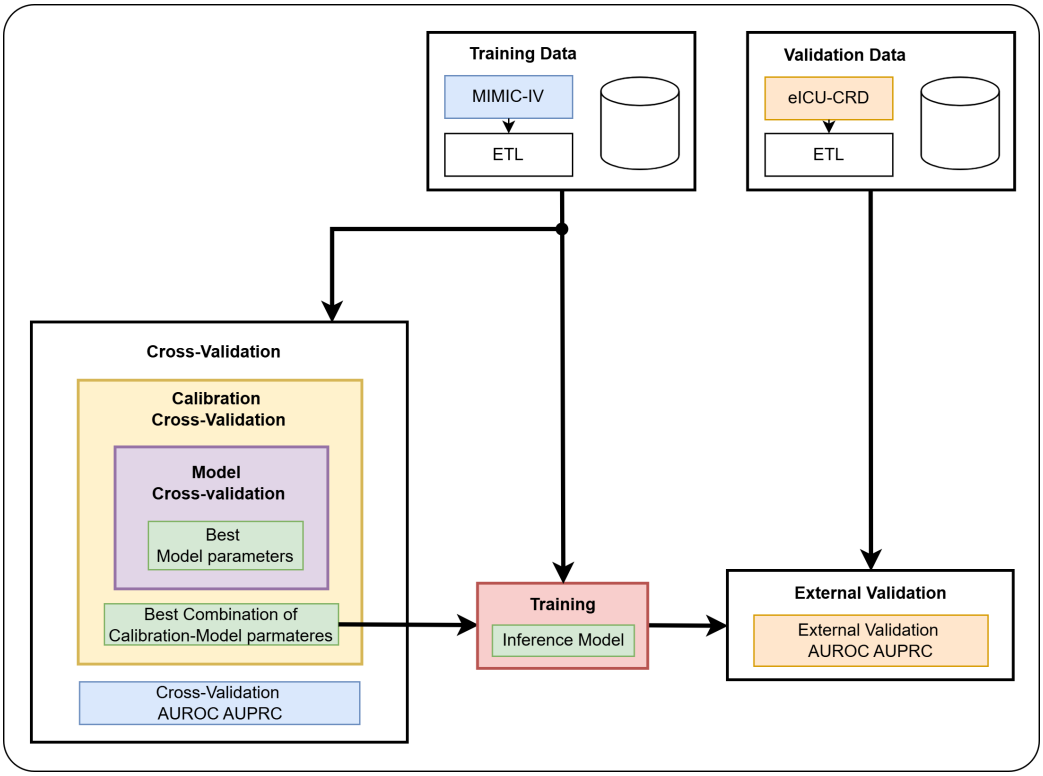
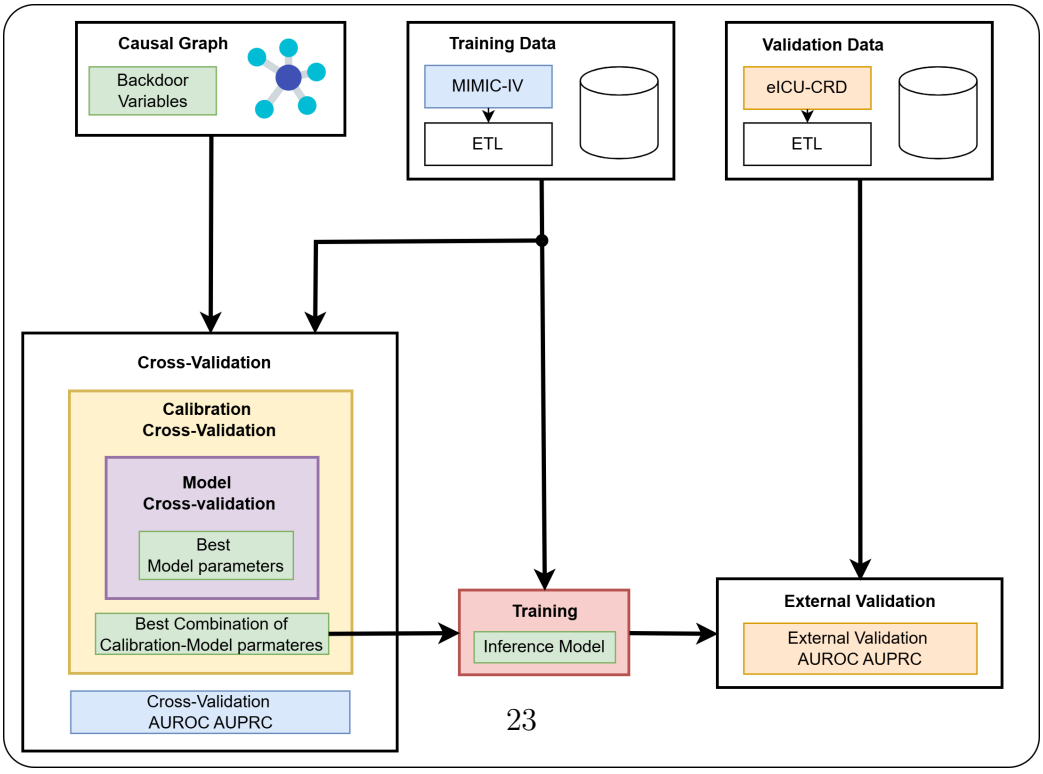


Figure S3: Formal causal assumptions for our study. Treatments N, F are represented in bold blue, backdoors $\mathbf{Z}$ and measurement noise $\mathbf{W}$ in dashed black, outcomes Y in bold red.



(a)



(b)

Figure S4: Training pipelines for the (S4a) Predictive model and (S4b) Causal model

Feature Name	s3	s4	s5	s30	s31	s32	s33	s34	s35
WBC [K/uL]	8.4	14.2	14.2	16.1	26.0	12.2	26.0	16.1	10.4
Residual SOFA	3.7	3.7	1.0	0.0	3.7	3.7	0.8	1.7	3.7
Temperature [C]	36.9	37.6	35.9	36.5	36.4	36.4	36.4	37.2	36.4
Urine Output [ml/kg/hr]	0.5	2.4	0.4	0.1	0.9	1.2	1.6	1.2	0.2
Mean Blood Pressure [mmHg]	64.7	54.3	73.1	54.3	64.7	68.0	73.1	73.1	64.7
Lactate [mmol/L]	3.4	3.4	3.8	3.8	9.4	2.9	5.4	6.6	9.4
Admission Age [Year(s)]	66.0	70.0	56.0	56.0	84.0	61.0	66.0	70.0	79.0
Diastolic Blood Pressure [mmHg]	44.0	46.0	46.0	39.0	41.0	60.0	44.0	44.0	49.0
Creatinine [mg/dL]	1.0	0.8	2.1	2.1	1.7	1.0	1.7	2.7	1.1
Heart Rate [BPM]	66.3	97.2	75.2	66.3	85.6	80.7	75.2	85.6	66.3

(a) Clinical baselines of patients in Fig. (5a, 5b, 5c) when evaluating the effect of intravenous fluids treatment on clinical Improvement endpoint.

Feature Name	s42	s43	s44	s45	s46	s47	s60	s61	s62
WBC [K/uL]	6.1	26.0	6.1	6.1	14.2	14.2	8.4	26.0	16.1
Residual SOFA	2.8	3.7	2.8	3.7	1.7	2.8	3.7	3.7	3.7
Temperature [C]	36.8	37.0	37.2	36.8	37.0	36.9	37.6	36.4	36.4
Urine Output [ml/kg/hr]	0.4	2.4	0.7	0.1	2.4	0.9	0.2	0.9	0.9
Mean Blood Pressure [mmHg]	64.7	68.0	68.0	54.3	63.3	57.0	68.0	64.7	64.7
Lactate [mmol/L]	2.2	2.2	3.8	9.4	3.8	9.4	2.2	9.4	9.4
Admission Age [Year(s)]	36.0	49.0	74.0	79.0	66.0	84.0	66.0	84.0	84.0
Diastolic Blood Pressure [mmHg]	51.0	55.0	41.0	39.0	46.0	44.0	60.0	41.0	44.0
Creatinine [mg/dL]	2.1	0.8	1.1	1.3	0.8	1.7	1.4	1.7	0.8
Heart Rate [BPM]	109.8	118.1	75.2	85.6	103.4	85.6	103.4	85.6	103.4

(b) Clinical baselines of patients in Fig. (5d,5e,5f) when evaluating the effect of norepinephrine equivalent dose treatment on clinical Improvement endpoint.

Feature Name	s93	s94	s95	s186	s187	s188	s207	s208	s209
WBC [K/uL]	18.2	26.0	14.2	12.2	8.4	18.2	16.1	14.2	16.1
Residual SOFA	0.3	0.0	0.0	3.7	1.7	3.7	1.7	0.0	1.7
Temperature [C]	37.0	36.4	36.4	37.6	36.9	36.9	36.4	36.7	37.2
Urine Output [ml/kg/hr]	1.6	0.2	1.6	0.1	0.2	1.2	1.2	1.2	0.1
Mean Blood Pressure [mmHg]	73.1	54.3	59.3	59.3	59.3	68.0	57.0	57.0	64.7
Lactate [mmol/L]	6.6	3.4	9.4	6.6	2.5	9.4	2.9	9.4	2.5
Admission Age [Year(s)]	70.0	49.0	36.0	61.0	84.0	66.0	70.0	56.0	49.0
Diastolic Blood Pressure [mmHg]	44.0	41.0	41.0	46.0	46.0	49.0	51.0	39.0	44.0
Creatinine [mg/dL]	1.1	1.3	1.4	1.4	1.0	2.1	1.1	4.1	0.8
Heart Rate [BPM]	109.8	80.7	66.3	75.2	109.8	118.1	92.0	118.1	103.4

(c) Clinical baselines of patients in Fig. (5g,5h,5i) when evaluating the effect of intravenous fluids treatment on in-hospital survival.

Feature Name	s261	s262	s263	s306	s307	s308	s363	s364	s365
WBC [K/uL]	8.4	12.2	16.1	14.2	6.1	12.2	10.4	14.2	21.0
Residual SOFA	0.8	3.7	0.3	0.0	2.8	3.7	0.0	1.0	1.7
Temperature [C]	36.4	36.4	37.0	36.7	36.8	36.4	36.7	35.9	36.5
Urine Output [ml/kg/hr]	0.5	1.2	0.7	0.4	0.7	0.9	1.2	0.9	2.4
Mean Blood Pressure [mmHg]	59.3	68.0	61.3	68.0	54.3	54.3	68.0	73.1	57.0
Lactate [mmol/L]	6.6	2.9	2.2	3.8	3.4	9.4	2.5	6.6	9.4

D. Reporting Guidelines

D.1. TRIPOD+AI

Item No.	Item	Section(s)
1	Identify the study as developing or evaluating the performance of a multivariable prediction model, the target population, and the outcome to be predicted	1
2	See TRIPOD+AI for Abstracts checklist	
3a	Explain the healthcare context (including whether diagnostic or prognostic) and rationale for developing or evaluating the prediction model, including references to existing models	1
3b	Describe the target population and the intended purpose of the prediction model in the context of the care pathway, including its intended users (e.g., healthcare professionals, patients, public)	1
3c	Describe any known health inequalities between sociodemographic groups	2.5
4	Specify the study objectives, including whether the study describes the development or validation of a prediction model (or both)	1, 2.1

Item No.	Item	Section(s)
5a	Describe the sources of data separately for the development and evaluation datasets (e.g., randomised trial, cohort, routine care or registry data), the rationale for using these data, and representativeness of the data	2.3
5b	Specify the dates of the collected participant data, including start and end of participant accrual; and, if applicable, end of follow-up	2.3
6a	Specify key elements of the study setting (e.g., primary care, secondary care, general population) including the number and location of centres	2.3
6b	Describe the eligibility criteria for study participants	2.3
6c	Give details of any treatments received, and how they were handled during model development or evaluation, if relevant	2.3, 2.4, 2.5
7	Describe any data pre-processing and quality checking, including whether this was similar across relevant sociodemographic groups	2.4
8a	Clearly define the outcome that is being predicted and the time horizon, including how and when assessed, the rationale for choosing this outcome, and whether the method of outcome assessment is consistent across sociodemographic groups	2.4

Item No.	Item	Section(s)
8b	If outcome assessment requires subjective interpretation, describe the qualifications and demographic characteristics of the outcome assessors	
8c	Report any actions to blind assessment of the outcome to be predicted	
9a	Describe the choice of initial predictors (e.g., literature, previous models, all available predictors) and any pre-selection of predictors before model building	2.2
9b	Clearly define all predictors, including how and when they were measured (and any actions to blind assessment of predictors for the outcome and other predictors)	2.4, Supplementary B
9c	If predictor measurement requires subjective interpretation, describe the qualifications and demographic characteristics of the predictor assessors	
10	Explain how the study size was arrived at (separately for development and evaluation), and justify that the study size was sufficient to answer the research question. Include details of any sample size calculation	2.3
11	Describe how missing data were handled. Provide reasons for omitting any data	2.4, 2.5, Supplementary B

Item No.	Item	Section(s)
12a	Describe how the data were used (e.g., for development and evaluation of model performance) in the analysis, including whether the data were partitioned, considering any sample size requirements	2.5
12b	Depending on the type of model, describe how predictors were handled in the analyses (functional form, rescaling, transformation, or any standardisation).	2.4, 2.5
12c	Specify the type of model, rationale, all model-building steps, including any hyperparameter tuning, and method for internal validation	2.5
12d	Describe if and how any heterogeneity in estimates of model parameter values and model performance was handled and quantified across clusters (e.g., hospitals, countries). See TRIPOD-Cluster for additional considerations	
12e	Specify all measures and plots used (and their rationale) to evaluate model performance (e.g., discrimination, calibration, clinical utility) and, if relevant, to compare multiple models	2.5
12f	Describe any model updating (e.g., recalibration) arising from the model evaluation, either overall or for particular sociodemographic groups or settings	

Item No.	Item	Section(s)
12g	For model evaluation, describe how the model predictions were calculated (e.g., formula, code, object, application programming interface)	2.5, 2.6
13	If class imbalance methods were used, state why and how this was done, and any subsequent methods to recalibrate the model or the model predictions	
14	Describe any approaches that were used to address model fairness and their rationale	2.5
15	Specify the output of the prediction model (e.g., probabilities, classification). Provide details and rationale for any classification and how the thresholds were identified	2.6
16	Identify any differences between the development and evaluation data in healthcare setting, eligibility criteria, outcome, and predictors	3.1, Supplementary B
17	Name the institutional research board or ethics committee that approved the study and describe the participant-informed consent or the ethics committee waiver of informed consent	9.2
18a	Give the source of funding and the role of the funders for the present study	
18b	Declare any conflicts of interest and financial disclosures for all authors	9.1

Item No.	Item	Section(s)
18c	Indicate where the study protocol can be accessed or state that a protocol was not prepared	
18d	Provide registration information for the study, including register name and registration number, or state that the study was not registered	
18e	Provide details of the availability of the study data	6
18f	Provide details of the availability of the analytical code	7
19	Provide details of any patient and public involvement during the design, conduct, reporting, interpretation, or dissemination of the study or state no involvement.	
20a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.	3.1, Supplementary B.4

Item No.	Item	Section(s)
20b	Report the characteristics overall and, where applicable, for each data source or setting, including the key dates, key predictors (including demographics), treatments received, sample size, number of outcome events, follow-up time, and amount of missing data. A table may be helpful. Report any differences across key demographic groups.	3.1, Supplementary B
20c	For model evaluation, show a comparison with the development data of the distribution of important predictors (demographics, predictors, and outcome).	Supplementary B
21	Specify the number of participants and outcome events in each analysis (e.g., for model development, hyperparameter tuning, model evaluation)	2.5, 3.1, Supplementary B.2
22	Provide details of the full prediction model (e.g., formula, code, object, application programming interface) to allow predictions in new individuals and to enable third-party evaluation and implementation, including any restrictions to access or re-use (e.g., freely available, proprietary)	7

Item No.	Item	Section(s)
23a	Report model performance estimates with confidence intervals, including for any key subgroups (e.g., sociodemographic). Consider plots to aid presentation.	3.3, Supplementary B.2
23b	If examined, report results of any heterogeneity in model performance across clusters. See TRIPOD Cluster for additional details.	3.4
24	Report the results from any model updating, including the updated model and subsequent performance	
25	Give an overall interpretation of the main results, including issues of fairness in the context of the objectives and previous studies	4
26	Discuss any limitations of the study (such as a non-representative sample, sample size, overfitting, missing data) and their effects on any biases, statistical uncertainty, and generalizability	4.1
27a	Describe how poor quality or unavailable input data (e.g., predictor values) should be assessed and handled when implementing the prediction model	4.1

Item No.	Item	Section(s)
27b	Specify whether users will be required to interact in the handling of the input data or use of the model, and what level of expertise is required of users	
27c	Discuss any next steps for future research, with a specific view to applicability and generalizability of the model	4

Table S10: TRIPOD+AI Checklist[4]

D.2. TARGET Statement

Item No.	Item	Section(s)
1a	Identify that the study attempts to emulate a target trial using observational data. State the study objectives and briefly summarize the specified target trial.	1
1b	Report the data sources used for emulation.	1
1c	Summarize key assumptions, statistical methods, findings and conclusions.	1
2	Describe the scientific background of the study and the gap in knowledge.	1
3	Summarize the causal question.	2.1, 3.2
4	Describe the rationale for emulating a target trial with the available data. Cite randomized trials informing the design of the target trial if applicable.	1

Item No.	Item	Section(s)
5	Cite the data sources contributing to the analyses and for each one describe the following: original purpose, type, the geographic locations, setting and time-period. If relevant, describe how the data were linked or pooled.	2.3
6a	Target trial specification: Describe the eligibility criteria.	2.3, 3.1
7a	Target trial emulation: Describe how the eligibility criteria were operationalized with the data.	2.3, 2.4, Supplementary B
6b	Target trial specification: Describe the treatment strategies that would be compared.	2.2, 3.1
7b	Target trial emulation: Describe how the treatment strategies were operationalized with the data.	2.4, 3.1, Supplementary B
6c	Target trial specification: Report that eligible individuals would be randomly assigned to treatment strategies and may be aware of their treatment allocation.	2.1, 3.3
7c	Target trial emulation: Describe how assignment to treatment strategies was operationalized with the data.	2.1, 2.6, 3.3
6d	Target trial specification: Clarify that follow-up would start at time of assignment to the treatment strategies. Specify when follow-up would end.	2.3, 3.1

Item No.	Item	Section(s)
7d	Target trial emulation: Clarify that follow-up starts at the time individuals were assigned to the treatment strategies. Describe how the end of follow-up was operationalized with the data.	2.3, 2.4
6e	Target trial specification: Describe the outcomes.	2.4, 3.1
7e	Target trial emulation: Describe how the outcomes were operationalized with the data.	2.4, Supplementary B
6f	Target trial specification: Describe the causal contrasts of interest, including effect measures.	2.1, 3.3
7f	Target trial emulation: Describe how the causal contrasts were operationalized with the data, including effect measures.	2.5, 2.6, 3.3
6g	Target trial specification: Describe assumptions that would be made to identify each causal estimand. Describe the variables, if any, related to these assumptions.	2.1, 2.2, 3.3
7g.i	Target trial emulation: For each causal estimand, describe assumptions made to identify it, including assumptions regarding baseline confounding due to lack of randomization.	2.1, 2.2, 3.3
7g.ii	Target trial emulation: Describe how the variables related to these assumptions were operationalized with the data.	2.3, 2.4, Supplementary B

Item No.	Item	Section(s)
6h	Target trial specification: For each causal estimand, describe the data analysis procedures and any associated statistical modelling assumptions, including approaches for handling missing data.	2.5, 2.6
7h.i	Target trial emulation: For each causal estimand, describe the data analysis procedures and any associated statistical modelling assumptions, including approaches for handling missing data.	2.5, 2.6
7h.ii	Target trial emulation: For each causal estimand, describe any additional analyses conducted to assess the sensitivity of the results to the choice of operationalizations, assumptions and analysis.	2.7, 3.5, Supplementary B.3
8	Report numbers of individuals assessed for eligibility, eligible, and assigned to each treatment strategy. A flow diagram is strongly recommended.	3.1, Supplementary B
9	Describe the distribution of characteristics of individuals at baseline, by treatment strategy.	3.1, Supplementary B
10	Summarize length of follow-up and describe reasons for end of follow-up for each treatment strategy and causal contrast.	2.3, 3.1
11	Describe the frequency of missing data in all variables, by treatment strategy when applicable.	Supplementary B
12	Describe the frequency or distribution of each outcome, by treatment strategy.	3.1, Supplementary B

Item No.	Item	Section(s)
13	Report the effect estimates for each causal contrast with corresponding measures of precision, including both absolute and relative measures of effect, when applicable.	3.3, 3.4, Supplementary B.3
14	Report results of all analyses to assess the sensitivity of the estimates to choices in operationalizations, assumptions and analysis.	3.5, Supplementary B.3
15	Provide an interpretation of the key findings.	4
16	Discuss the limitations of the study considering differences between the target trial and its emulation and the plausibility of assumptions, including assumptions regarding baseline confounding due to lack of randomization.	4.1
17	Provide the institutional research board or ethics committee that approved the study and approval numbers, if relevant.	9.2
18	State whether, when and where the study protocol was registered.	
19	Provide information on whether data, analytic code and/or other materials are accessible, and where and how they can be accessed.	6, 7
20	Provide the sources of funding and detail the role of the funders in the design, conduct and reporting of the study.	

Item No.	Item	Section(s)
21	State any conflicts of interest and financial disclosures for all authors.	9.1

Table S11: Target Statement Checklist[3]