Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

# Step-by-step causal analysis of EHRs to ground decision-making

**Matthieu Doutreligne[1,2*], Tristan Struja[3,4], Judith Abecassis[1], Claire Morgand[5], Leo Anthony Celi[3,6,7], Gaël Varoquaux[1]**

**Tanawat Wuttiyakorn**
**M.Sc. Student**
**Department of Clinical Epidemiology and Biostatistics, Faculty of Medicine Ramathibodi Hospital, Mahidol university**

# Outline

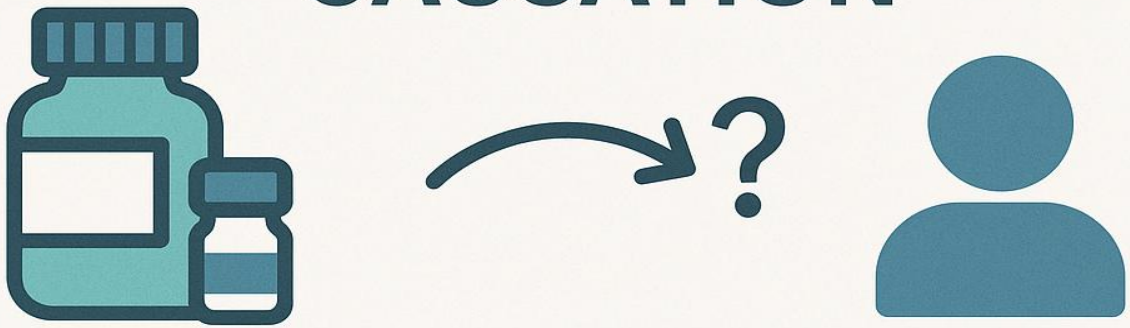
- Pitfalls of Observational Data
- 5-step analytic framework
- Application

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

# PREDICTION

X predicts Y

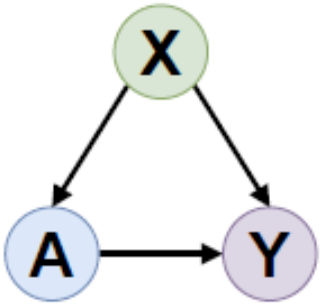
# CAUSATION



Does T cause Y?

# Pitfalls of Observational Data

- **Confounding**: Common causes affecting both treatment choice and outcome.



Confounder

- **Selection Bias**: Groups being compared differ systematically due to selection, not just treatment.

# Pitfalls of Observational Data

- **Immortal Time Bias**

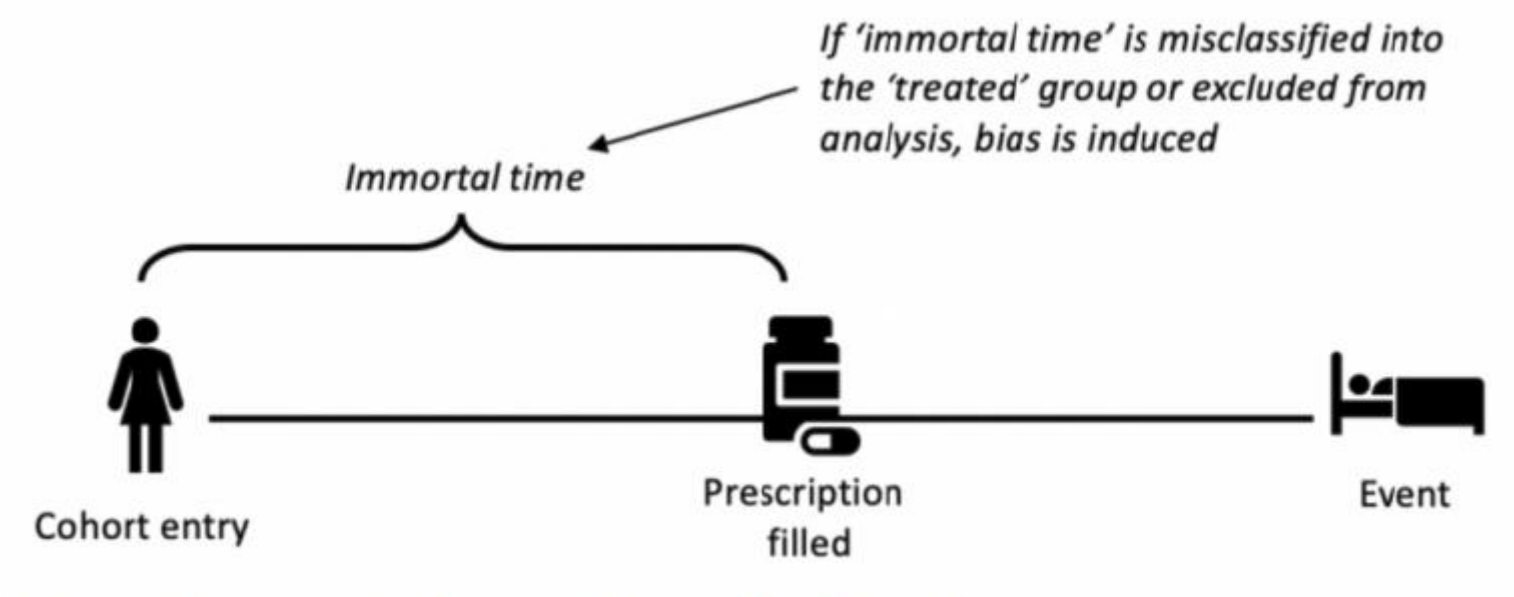
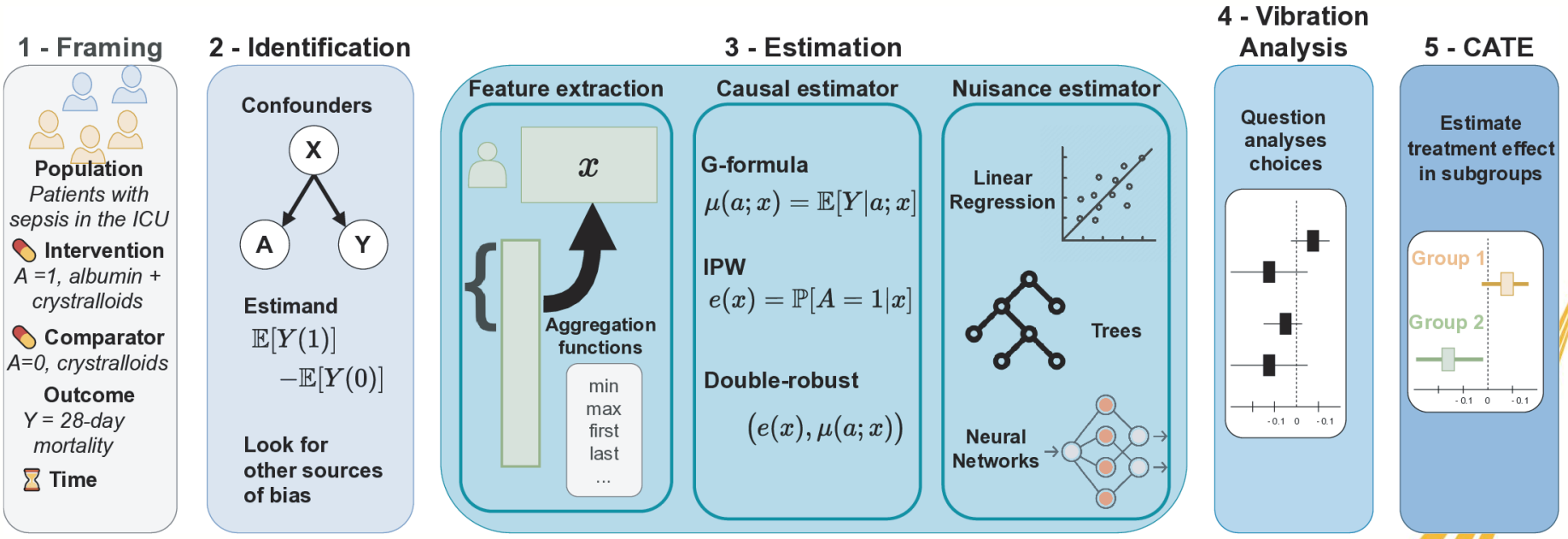



Fig 1. Immortal time bias illustration.

# 5-step analytic framework



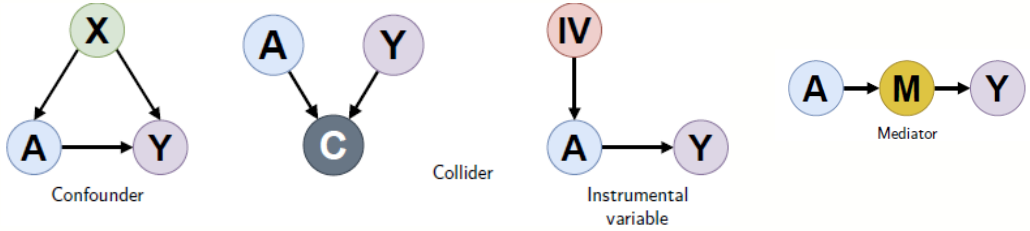
# Step 1 - Study Design

**Table 1. PICO(T) components help to clearly define the medical question of interest.**

| PICO component | Description | Notation | Example |
|---|---|---|---|
| **Population** | What is the target population of interest? | $X \sim \mathbb{P}(X)$, the covariate distribution | Patients with sepsis in the ICU |
| **Intervention** | What is the treatment? | $A \sim \mathbb{P}(A = 1) = p_A$, the probability to be treated | Combination of crystalloids and albumin |
| **Control** | What is the clinically relevant comparator? | $1 - A \sim 1 - p_A$ | Crystalloids only |
| **Outcome** | What are the outcomes to compare? | $Y(1), Y(0) \sim \mathbb{P}(Y(1), Y(0))$, the potential outcomes distribution | 28-day mortality |
| **Time** | Is the start of follow-up aligned with intervention assignment? | N/A | Intervention within the first day |

# Step 2 - Identification

1. Stating Causal Assumptions
   - e.g., Unconfoundedness, Overlap, No Interference
2. Categorizing Covariates



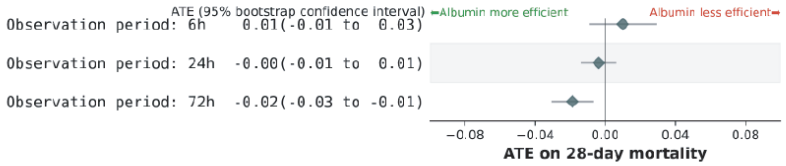

3. Using Directed Acyclic Graphs (DAGs)
4. Defining the Estimand
5. Choosing Causal Estimators
   - e.g., G-formula, PSM, IPW, DML

# Step 3 - Statistical estimation

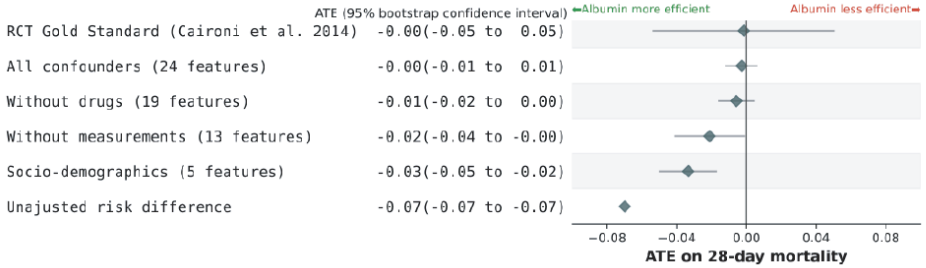
- Confounder Aggregation
- Missing Value Handling
- Outcome and Treatment Estimators
- Hyperparameter Tuning

# Step 4 - Vibration analysis

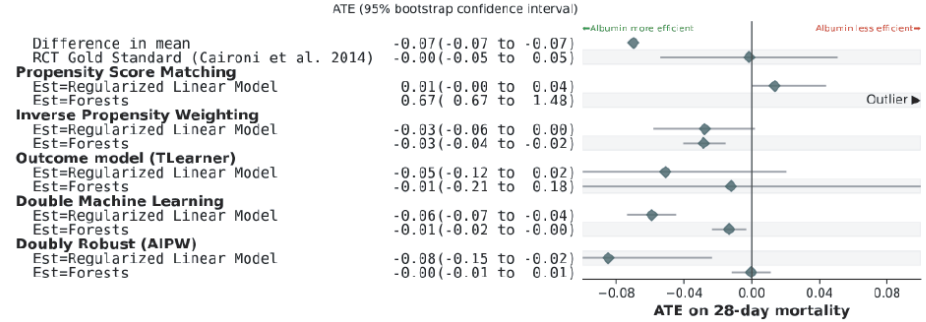
(a) *Framing – Immortal Time Bias*



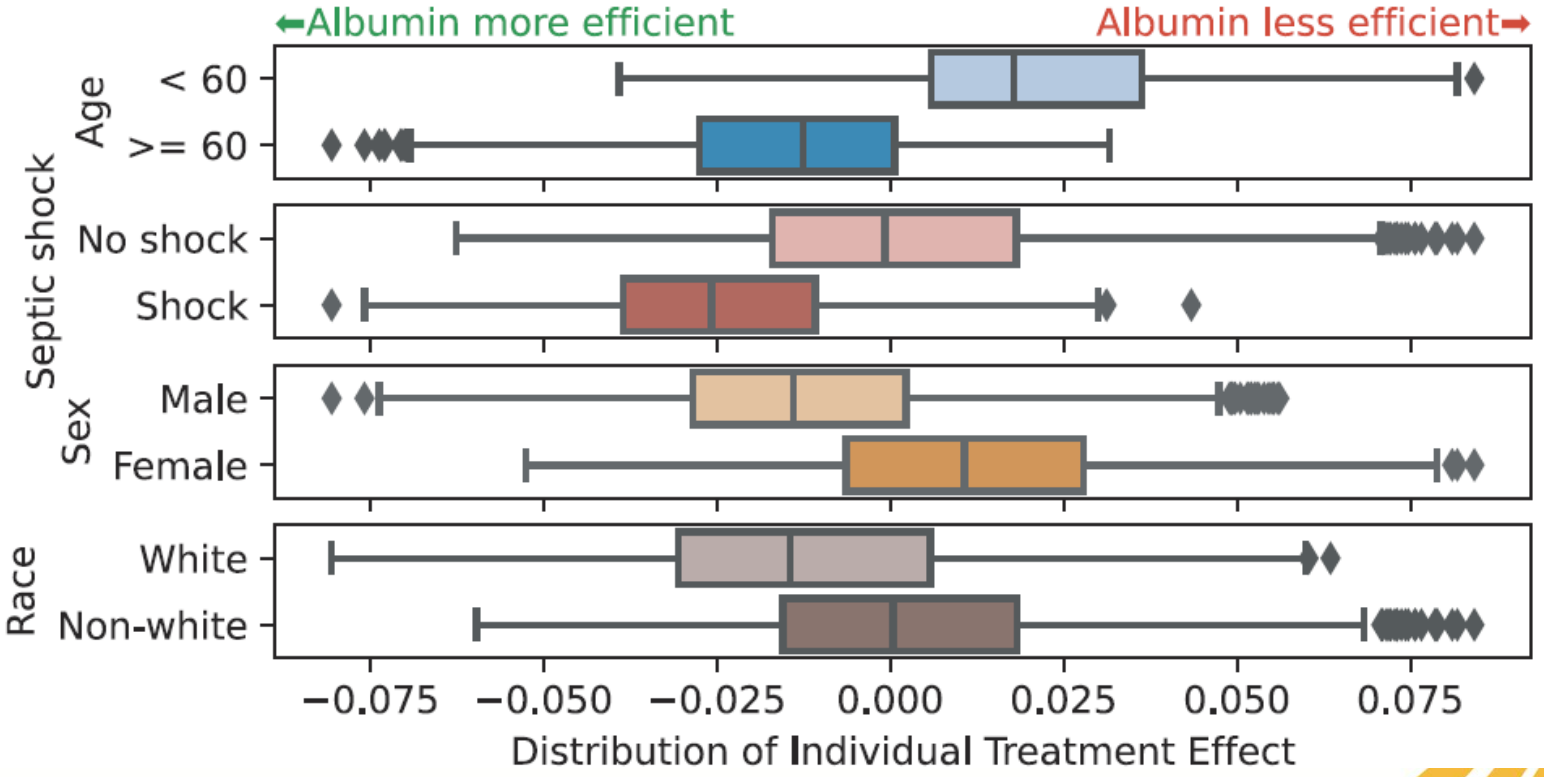

(b) *Identification – confounders choice*




(c) *Model selection*



# Step 5 - Treatment heterogeneity



# Application: Albumin vs. Crystalloids in Sepsis

- Patients with septic shock need fluids to maintain blood flow to organs
  - Crystalloids (saline) are common, cheap, and safe, but much of the fluid leaks out of the blood vessels
  - Colloids (albumin) are thought to stay in the blood vessels better, potentially improving circulation more effectively, but they are more expensive and might have adverse effects.

- **Clinical Question**: What is the effect of using albumin combined with crystalloids compared to using crystalloids alone on 28-day mortality in sepsis patients?
- **Data**: MIMIC-IV ICU database.
- **Validation Strategy**: Compare estimated ATE to known RCT average null effect.

# Step 1 - Study Design

**P**opulation :

- Patients identified with sepsis during an ICU stay, >= 18 years old, and at least 24 hours of follow-up data available.

**I**ntervention:

- Receiving a combination of crystalloids and albumin within the first 24 hours of the ICU stay.

**C**ontrol:

- Receiving crystalloids only within the first 24 hours of the ICU stay.

**O**utcome:

- Death within 28 days

# Step 1 - Study Design

Time :

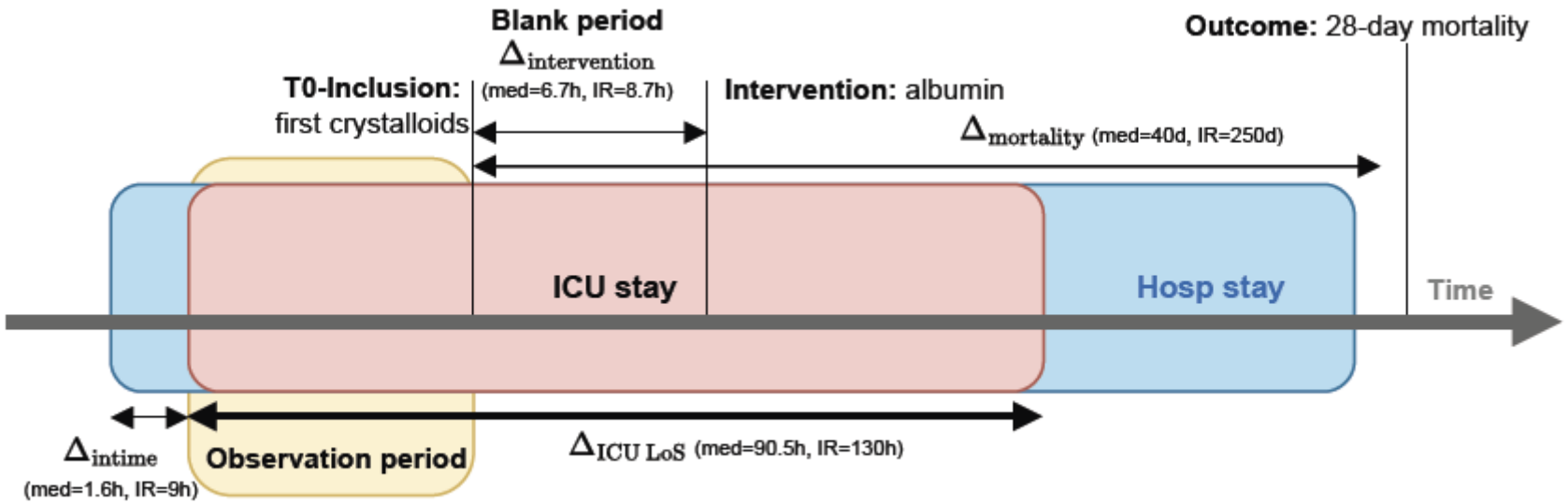
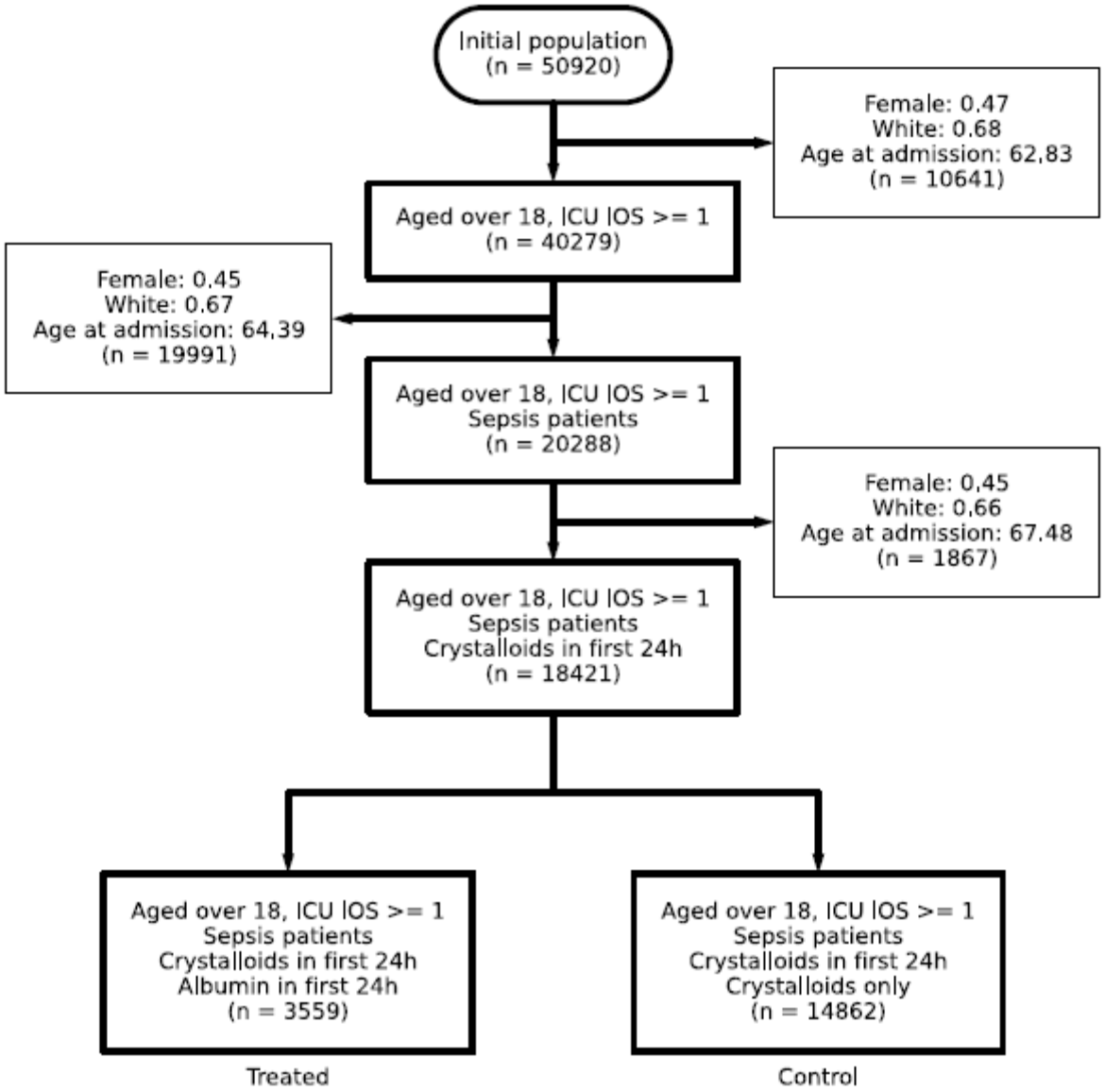



Fig 1. Graphical timeline.

# Selection flowchart.



| | Missing | Overall | Cristalloids only | Cristalloids + Albumin | P-Value |
|---|---|---|---|---|---|
| n | | 18421 | 14862 | 3559 | |
| Glycopeptide, n (%) | | 9492 (51.5) | 7650 (51.5) | 1842 (51.8) | |
| Beta-lactams, n (%) | | 5761 (31.3) | 5271 (35.5) | 490 (13.8) | |
| Carbapenems, n (%) | | 727 (3.9) | 636 (4.3) | 91 (2.6) | |
| Aminoglycosides, n (%) | | 314 (1.7) | 290 (2.0) | 24 (0.7) | |
| suspected_infection_blood, n (%) | | 170 (0.9) | 149 (1.0) | 21 (0.6) | |
| RRT, n (%) | | 229 (1.2) | 205 (1.4) | 24 (0.7) | |
| ventilation, n (%) | | 16376 (88.9) | 12931 (87.0) | 3445 (96.8) | |
| vasopressors, n (%) | | 9058 (49.2) | 6204 (41.7) | 2854 (80.2) | |
| Female, n (%) | | 7653 (41.5) | 6322 (42.5) | 1331 (37.4) | |
| White, n (%) | | 12366 (67.1) | 9808 (66.0) | 2558 (71.9) | |
| Emergency admission, n (%) | | 9605 (52.1) | 8512 (57.3) | 1093 (30.7) | |
| Insurance, Medicare, n (%) | | 9727 (52.8) | 7958 (53.5) | 1769 (49.7) | |
| myocardial_infarct, n (%) | | 3135 (17.0) | 2492 (16.8) | 643 (18.1) | |
| malignant_cancer, n (%) | | 2465 (13.4) | 2128 (14.3) | 337 (9.5) | |
| diabetes_with_cc, n (%) | | 1633 (8.9) | 1362 (9.2) | 271 (7.6) | |
| diabetes_without_cc, n (%) | | 4369 (23.7) | 3532 (23.8) | 837 (23.5) | |
| metastatic_solid_tumor, n (%) | | 1127 (6.1) | 1016 (6.8) | 111 (3.1) | |
| severe_liver_disease, n (%) | | 1289 (7.0) | 880 (5.9) | 409 (11.5) | |
| renal_disease, n (%) | | 3765 (20.4) | 3159 (21.3) | 606 (17.0) | |
| aki_stage_0.0, n (%) | | 7368 (40.0) | 6284 (42.3) | 1084 (30.5) | |
| aki_stage_1.0, n (%) | | 4019 (21.8) | 3222 (21.7) | 797 (22.4) | |
| aki_stage_2.0, n (%) | | 6087 (33.0) | 4605 (31.0) | 1482 (41.6) | |
| aki_stage_3.0, n (%) | | 947 (5.1) | 751 (5.1) | 196 (5.5) | |
| SOFA, mean (SD) | 0 | 6.0 (3.5) | 5.7 (3.4) | 6.9 (3.6) | <0.001 |
| SAPSII, mean (SD) | 0 | 40.3 (14.1) | 39.8 (14.1) | 42.8 (13.6) | <0.001 |
| Weight, mean (SD) | 97 | 83.3 (23.7) | 82.5 (24.2) | 86.4 (21.2) | <0.001 |
| temperature, mean (SD) | 966 | 36.9 (0.6) | 36.9 (0.6) | 36.8 (0.6) | <0.001 |
| mbp, mean (SD) | 0 | 75.6 (10.2) | 76.3 (10.7) | 72.4 (7.2) | <0.001 |
| resp_rate, mean (SD) | 9 | 19.3 (4.3) | 19.6 (4.4) | 18.0 (3.8) | <0.001 |
| heart_rate, mean (SD) | 0 | 86.2 (16.3) | 86.2 (16.8) | 86.5 (14.3) | 0.197 |
| spo2, mean (SD) | 4 | 97.4 (2.2) | 97.3 (2.3) | 98.0 (2.1) | <0.001 |
| lactate, mean (SD) | 4616 | 3.0 (2.5) | 2.8 (2.4) | 3.7 (2.6) | <0.001 |
| urineoutput, mean (SD) | 301 | 24.0 (52.7) | 24.7 (58.2) | 21.1 (16.6) | <0.001 |
| admission_age, mean (SD) | 0 | 66.3 (16.2) | 66.1 (16.8) | 67.3 (13.1) | <0.001 |
| delta mortality to inclusion, mean (SD) | 11121 | 316.9 (640.2) | 309.6 (628.8) | 365.0 (708.9) | 0.022 |
| delta intervention to inclusion, mean (SD) | 14862 | 0.3 (0.2) | nan (nan) | 0.3 (0.2) | nan |
| delta inclusion to intime, mean (SD) | 0 | 0.1 (0.2) | 0.1 (0.2) | 0.1 (0.1) | 0.041 |
| delta ICU intime to hospital admission, mean (SD) | 0 | 1.1 (3.7) | 1.0 (3.7) | 1.6 (3.4) | <0.001 |
| los_hospital, mean (SD) | 0 | 12.6 (12.5) | 12.6 (12.5) | 12.9 (12.4) | 0.189 |
| los_icu, mean (SD) | 0 | 5.5 (6.7) | 5.5 (6.5) | 5.5 (7.2) | 0.605 |

Table 1. ***Characteristics of the trial population measured on the first 24 hours of ICU stay.***

# Step 2 - Identification

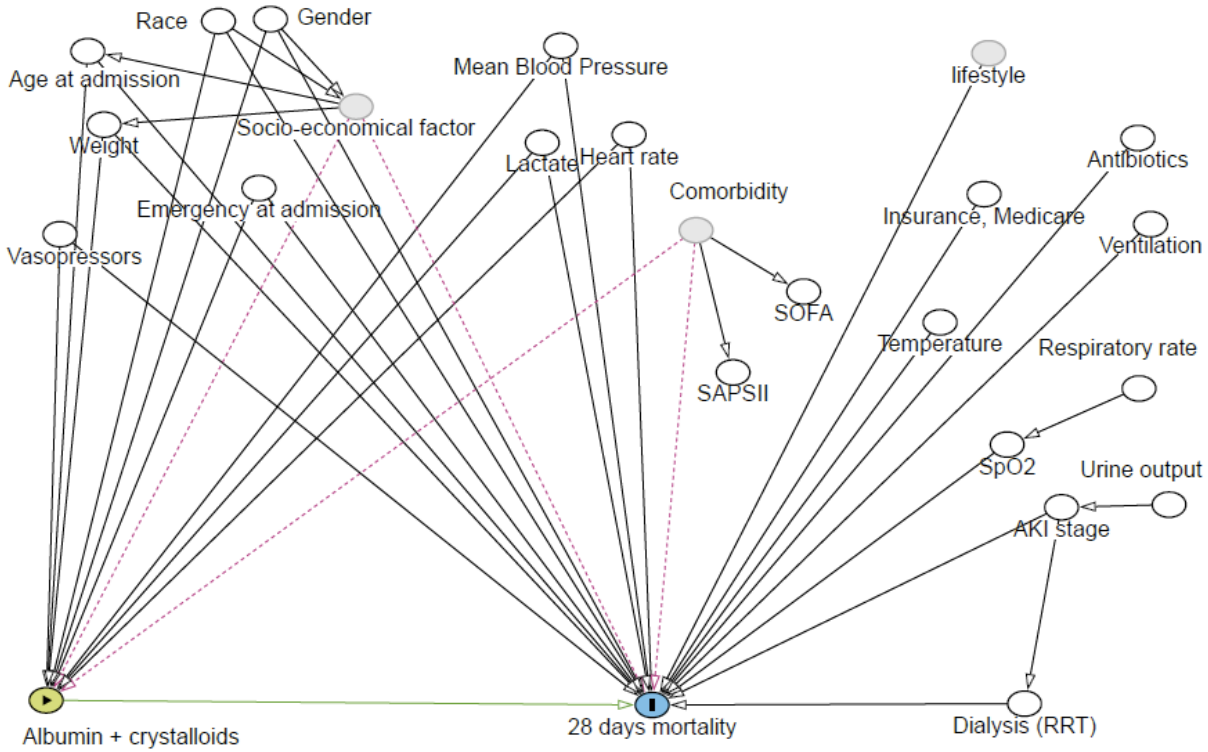



**Fig 1.** ***Causal graph for the Albumin vs crystalloids emulated trial***
*The green arrow indicates the effect studied. Black arrows show causal links known to medical expertise. Dotted red arrows highlight confounders not directly observed. For readability, we draw only the most important edges from an expert point of view. All white nodes correspond to variables included in our study.*

# Step 2 - Identification

**Causal estimation** strategies

- Inverse Propensity Weighting (IPW)
- Outcome modeling (G-formula)
- Augmented Inverse Propensity Weighting (AIPW)
- Double Machine Learning (DML)

**Software Implementation**:

| Packages | Simple installation | Confidence Intervals | sklearn estimator | sklearn pipeline | Propensity estimators | Doubly Robust estimators | TMLE estimator | Honest splitting (cross validation) |
|---|---|---|---|---|---|---|---|---|
| dowhy | ✓ | ✓ | ✓ | ✓ | ✓ | ✗ | ✗ | ✗ |
| EconML | ✓ | ✓ | ✓ | Yes except for imputers | ✗ | ✓ | ✗ | Only for doubly robust estimators |
| zEpid | ✓ | ✓ | ✗ | ✗ | ✓ | ✓ | ✓ | Only for TMLE |
| causalml | ✗ | ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | Only for doubly robust estimators |

Table 1. *Selection criteria for causal python packages.*

# Step 3 - Statistical estimation

Confounder Aggregation:

- Using the last recorded value before the follow-up period started.
- Using the first observed value.
- Using both the first and last values as two separate features

Missing Value Handling:

- Filled it in using the median value
- One-hot encoding

# Step 3 - Statistical estimation

Estimators:

- Main causal estimator: IPW, G-formula, AIPW, DML
- 2 Different types of machine learning models
  - Random Forests
  - Ridge Logistic Regression

Hyperparameter Tuning:

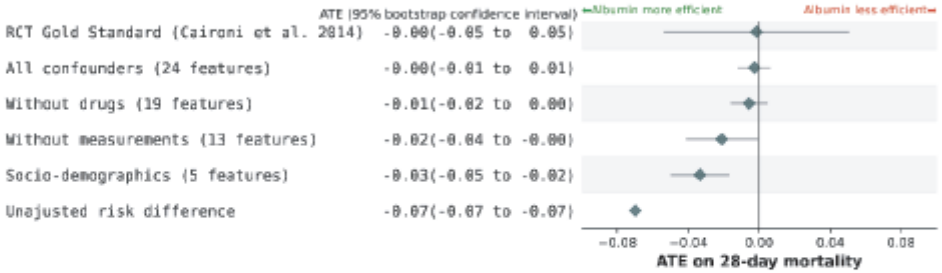
| Estimator type | estimator | nuisance | Grid |
|---|---|---|---|
| Linear | LogisticRegression | treatment | {'C': logspace(-3, 2, 10)} |
| Linear | Ridge | outcome | {'alpha': logspace(-3, 2, 10)} |
| Forest | RandomForestClassifier | treatment | {'n_estimators': ['10', '100', '200'], 'max_depth': ['3', '10', '50']} |
| Forest | RandomForestRegressor | outcome | {'n_estimators': ['10', '100', '200'], 'max_depth': ['3', '10', '50']} |

Table 1. ***Hyper-parameter grid used during random search optimization.***

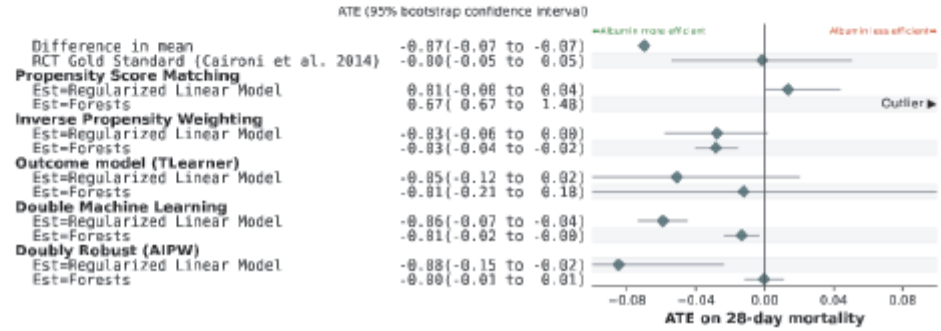
# Step 4 - Vibration analysis

(a) *Framing – Immortal Time Bias*


(b) *Identification – confounders choice*




(c) *Model selection*



# Step 5 - Treatment heterogeneity

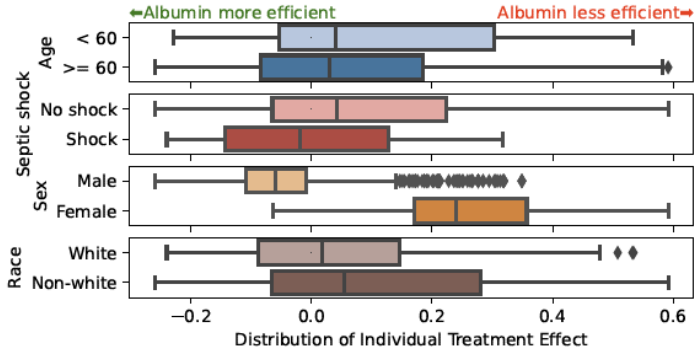



Fig 1. Values of Conditional Average Treatment effects on sex, age, race and pre-treatment septic shock estimated with a final forest estimator.
*The CATE are positive for each subgroups, which is not consistent with the null treatment effect obtained in the main analysis. The boxes contain between the 25th and 75th percentiles of the CATE distributions with the median indicated by a vertical line. The whiskers extends to 1.5 the inter-quartile range of the distribution.*

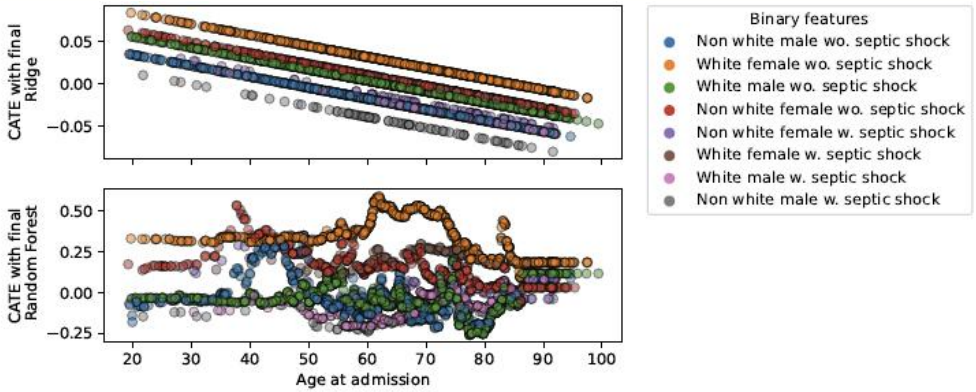



Fig 2. Values of Conditional Average Treatment effects on sex, age, race and pre-treatment septic shock plotted for different ages.

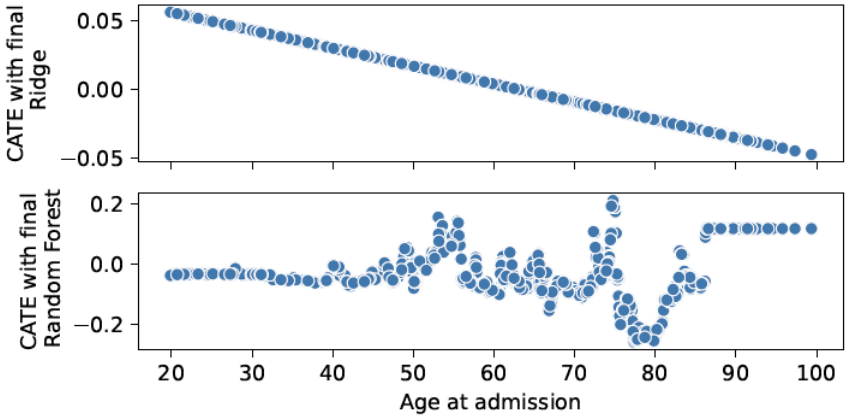



Fig 3. Values of Conditional Average Treatment effects on age, for the subpopulation of white male patients without septic shock.

Mahidol University
Faculty of Medicine Ramathibodi Hospital
Department of Clinical Epidemiology and Biostatistics

# Thank You!