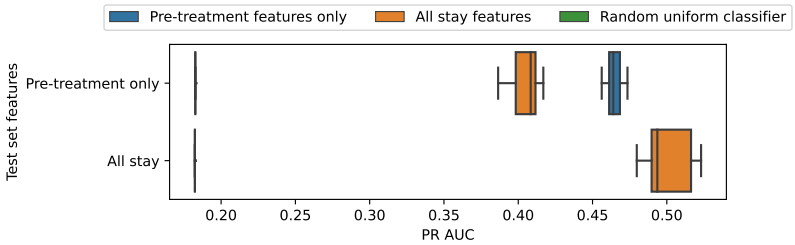
Supporting information

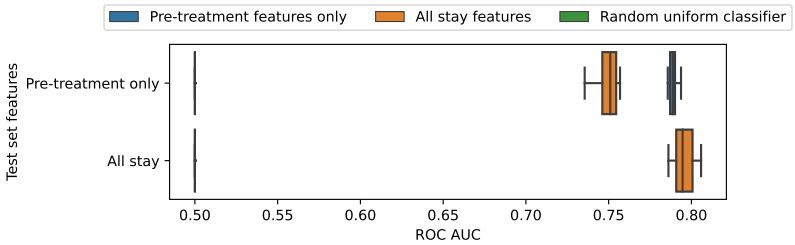
S1 Fig. Motivating example: Failure of predictive models to predict mortality from pretreatment variables. To illustrate how machine learning frameworks can fail to inform decision making, we present a motivating example from MIMIC-IV. Using the same population and covariates as in the main analysis (described in Table 6), we train a predictive model for 28-day mortality. We split the data into a training set (80%) and a test set (20%). The training set uses the last measurements from the first 24 hours, whereas the validation set only uses the last measurements before the administration of crystalloids. We split the train set into a train and a validation set. We fit a HistGradientBoosting classifier [1] on the train set and evaluate the performance on the validation set and on the test set. We see good area under the Precision-recall curve (PR AUC) on the validation set, but a deterioration of 10 points on the test set (Figure 5a). The same is seen in Figure 5b when measuring performances with Area Under the Curve of the Receiving Operator Characteristic (ROC AUC). On the contrary, a model trained on pre-treatment features yields competitive performances. This failure illustrates well the shortcuts on which predictive models could rely to make predictions. A clinically useful predictive model should support decision-making –in this case, addition of albumin to crystalloids– rather than maximizing predictive performance. In this example, causal thinking would have helped to identify the bias introduced by post-treatment features. In fact, these features should not be included in a causal analysis since they are post-treatment colliders.

This kind of error might sound naive to a clinical expert but relying on shortcuts –some of them being post-treatment variables– is a common error. Here, we detail some real use cases where machine learning fail in providing useful predictions for decision-making. [13] use deep learning to predict hip fracture using confounding patient and healthcare variables. An example of such covariates shown by the authors is the triage of patients before imaging that results in the model trying to predict the image acquisition machine and rely on it to predict hip fracture. [14] describe the use of algorithm in US extra-care programs. By equating care needs with previous care costs (in a pure predictive fashion), the algorithm falsely conclude that Black patients are healthier than equally white patients, since they do less money is spent on them for a given level of need. Beyond Machine Learning, we also spotted the inclusion of post-treatment variables in the development of the recent SCORE2 cardio-vascular risk score [106]: *Our risk models might have underestimated CVD risk be- cause data used to estimate multipliers were likely to include some people already on CVD prevention therapies (e.g. statins or anti- hypertensive medication.* This score might be used to inform on the initiation of statins for primary prevention. But, relying on post-treatment, it might under-discover patients who would benefit from statins at screening time.

[1] https://scikit-learn.org/stable/modules/ensemble.html#histogram-based-gradient-boosting



(a) *Area under the Precision-Recall curve (PR_AUC)*



(b) *Area under the Receiving Operator Characteristic (ROC_AUC)*

Fig 5. *Failure to predict 28-day mortality from a model fitted on pre-treatment variables. The model is trained on the last features from the whole stay and tested on two validation sets: one with all stay features and one with last features before crystalloids administration (Pre-treatment only). The all-stay model performance markedly decreases in the pre-treatment only dataset.*

S2 Fig. Immortal time bias illustration.

Figure 6 illustrates the immortal time bias. This time bias is a major pitfall in the retrospective evaluation of screening programs [107].

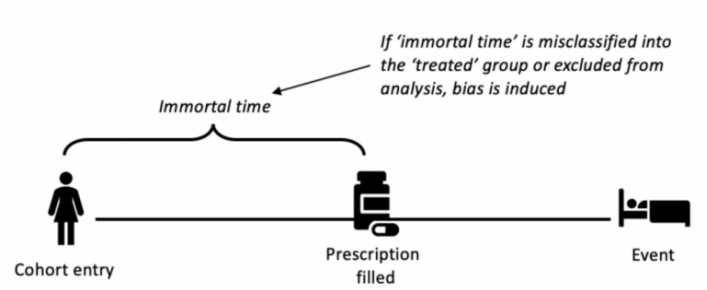


Fig 6. *Poor experimental design can introduce Immortal time bias, which leads to a treated group with falsely longer longevity [108].*

S3 Fig. Graphical timeline. Drawing a graphical timeline as the one in Figure 7 during the study design helps to detect and prevent time-related biases.

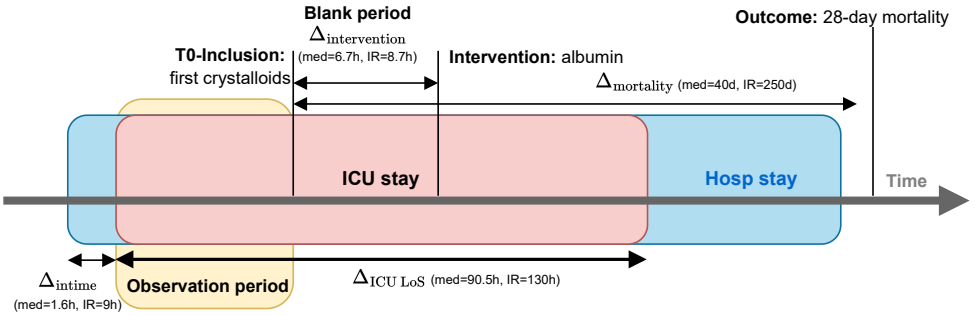


Fig 7. *Defining the inclusion event, the starting time T0 for follow-up, the intervention's assignment time and the observation window for confounders is crucial to avoid time and selection biases. In our study, the gap between the intervention and the inclusion is small compared to the occurrence of the outcome to limit immortal time bias: 6.7 hours vs 40 days for mortality.*

S4 Fig. Types of causal variables.

Figure 8 illustrates the different types of causal variables.

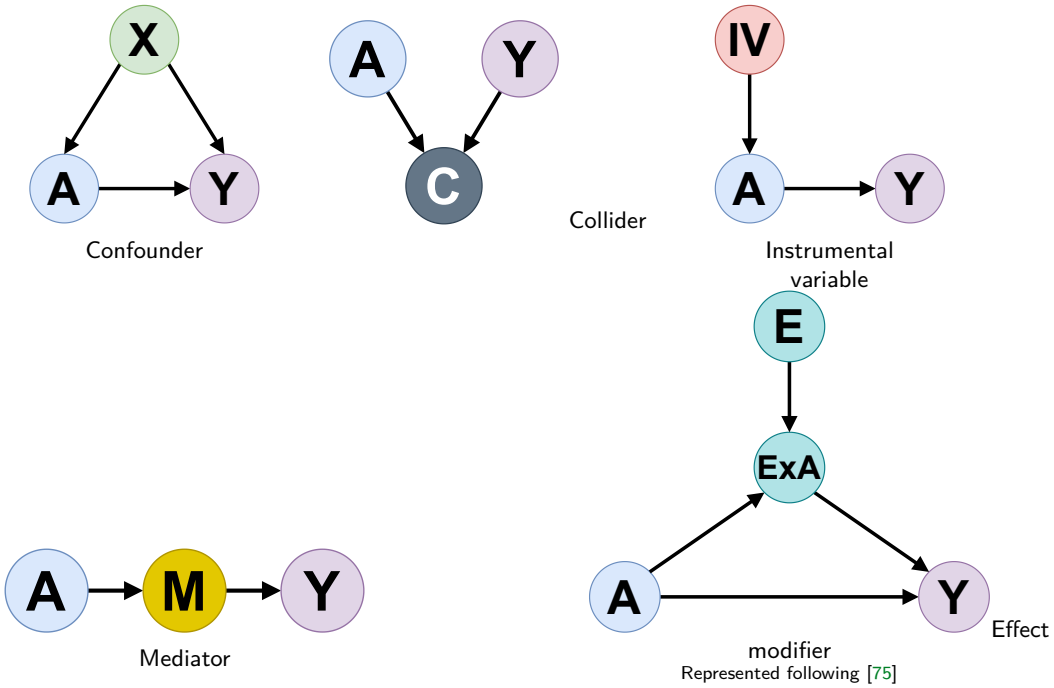


Fig 8. *The five categories of causal variables needed for our framework: A: Treatment, X: Confounder, IV: Instrumental variable, M: mediator, Y: Outcome, C: Collider, E: Effect modifier.*

S1 Appendix. Estimation of Treatment effect with MIMIC data.

We searched for causal inference studies in MIMIC using PubMed and Google scholar with the following search terms ((MIMIC-III OR MIMIC-IV) AND (causal inference OR treatment effect)). We retained eleven treatment effect studies clearly following the PICO framework:

- [109] studied the effect of High-flow nasal cannula oxygen (HFNC) against noninvasive mechanical ventilation on 801 patients with hypoxemia during ventilator weaning on 28-day mortality. They used propensity score matching, and found non-negative effects as previous RCTs reported –though those were focused on reintubation as the main outcome [110, 111].

- [112] studied the effect of lower hypoxemia vs higher hypoxemia thresholds for the initiation of invasive ventilation (defined with saturation-to-inspired oxygen ratio (SF)) for 3,357 patients from MIMIC receiving inspired oxygen fraction greater than 0.4 on 28-day mortality. Using bayesian G-computation (time-varying treatment model with gaussian process and outcome-model with BART, taking the treatment model as entry), they found protective effects for initialization at low hypoxemia. However, when externally validation their findings in the AmsterdamUMCdb dataset, they found the highest mortality probability for patients with low hypoxemia. Authors concluded that their model was heavily dependent on clinical context and baseline characteristics. There might be some starting-time bias in this study since it is really close

- [113] studied the effect of indwelling arterial catheters (IACs) vs non-IAC for 1,776 patients who are mechanically ventilated and did not require vasopressor support on 28-day mortality. They used propensity score matching and found no effect. A notebook based on google cloud access to MIMIC-IV replicating the study is available here.

- [114] studied the effect of transthoracic echocardiography vs no intervention for 6,361 patients with sepsis on 28-day mortality. They used IPW, PSM, g-formula and a doubly robust estimation. The propensity score was modeled with boosting and the outcome model with a logistic regression. They found a significant positive reduction of mortality (odd ratio 0.78, 95% CI 0.68-0.90). Study code is open source.

- [115] studied the effect of liberal –target SpO2 greater than 96%– vs conservative oxygenation –target SpO2 between 88-95%– in 4,062 mechanically ventilated patients on 90-day mortality. They found an advantage of the liberal strategy over liberal (ATE=0.13) by adjusting on age and apsii. This is not consistent with previous RCTs where no effects have been reported [116, 117].

- [118] studied the effect of fluid-limiting treatment –caped between 6 and 10 L– vs no cap on fluid administration strategies for 1,639 sepsis patients on 30 day-mortality. Using a dynamic Marginal Structural Model with IPW, they found a protective effect of fluid-limitation on ATE -0.01 (95%CI -0.016, -0.03). This is somehow concordant with the RIFTS RCT that found no effect of fluid limitation [119] and two previous meta-analyses [120, 121].

- [122] studied the effect of statin use prior to ICU admission vs absence of pre-ICU prescription for 8,200 patients with sepsis on 30-day mortality. Using AIPW (no estimator reported) and PSM (logistic regression), they found a decrease on mortality (ATE -0.039, 95%CI -0.084, -0.026). This partly supports previous findings in Propensity Matching bases observational studies [123, 124]. But all RCTs [125, 126] found no improvement for sepsis (not pre-admission administration though).

The [127] meta-analysis concludes that there is lack of evidence for the use of statins in sepsis with inconsistent results between RCTs (no effect) and observational studies (protective effect).

- [128] studied the effect of higher vs lower positive end-expiratory pressures (PEEP) in 1,411 patients with Acute Respiratory Distress Syndrome (ARDS) syndrome on 30 day mortality. Very few details on the methods were reported, but they found a protective effect for higher PEEP consistent results from a target trial [129].
- [128] also studied the effect of early use of a neuromuscular blocking agent vs placebo in 752 patients moderate-severe ARDS on 30 day mortality. Very few details on the methods were reported, but they found a protective effect for the use of a neuromuscular blocking agent, consistent with the results from a target trial [130].
- [95] studied the administration of a combination of albumin within the first 24-h after crystalloids vs crystalloids alone for 6,641 patients with sepsis on 28-day mortality. Using PSM, they found protective effect of combination on mortality, but insist on the importance of initialization timing. This is consistent with [92], who found a non-significant trend in favor of albumin used for severe sepsis patients and a significant reduction for septic shock patients, both on 90-day mortality. These results are aligned with [94] that found no effect for severe sepsis patient but positive effect for septic shock patients.
- [131] studied early enteral nutrition (EN) –<=53 ICU admission hours– vs delayed EN for 2,364 patients with sepsis and EN on acute kidney injury. With PSM, IPW and g-formula (logistic estimator each time), they found a protective effect (OR 0.319, 95%CI 0.245, 0.413) of EEN.

These eleven studies mainly used propensity score matching (6) and IPW (4), two of them used doubly robust methods, and only one included a non-linear estimator in either the outcome or the treatment model. None of them performed a vibration analysis on confounder selection or feature transformations. They have a strong focus on patients with sepsis. Only four of them found concordant results with previous RCTs [109, 118, 128].

S2 Appendix. Assumptions: what is needed for causal inference from observational studies.

The following four assumptions, referred as strong ignorability, are needed to assure identifiability of the causal estimands with observational data with most causal-inference methods [74], in particular these we use:

Assumption 1 (Unconfoundedness)

$$\{Y(0), Y(1)\} \perp\!\!\!\perp A|X \tag{1}$$

This condition –also called ignorability– is equivalent to the conditional independence on the propensity score $e(X) = \mathbb{P}(A = 1|X)$ *[132]:* $\{Y(0), Y(1)\} \perp\!\!\!\perp A|e(X)$.

Assumption 2 (Overlap, also known as Positivity)

$$\eta < e(x) < 1 - \eta \quad \forall x \in \mathcal{X} \text{ and some } \eta > 0 \tag{2}$$

The treatment is not perfectly predictable. Or in other words, every patient has a chance to be treated and not to be treated. For a given set of covariates, we need examples of both to recover the ATE.

As noted by [79], the choice of covariates X can be viewed as a trade-off between these two central assumptions. A bigger covariate set generally reinforces the ignorability assumption. In the contrary, overlap can be weakened by large $\mathcal{X}$ because of the potential inclusion of instrumental variables: variables only linked to the treatment which could lead to arbitrarily small propensity scores.

Assumption 3 (Consistency) *The observed outcome is the potential outcome of the assigned treatment:*

$$Y = A\,Y(1) + (1 - A)\,Y(0) \tag{3}$$

Here, we assume that the intervention A has been well defined. This assumption focuses on the design of the experiment. It clearly states the link between the observed outcome and the potential outcomes through the intervention [40].

Assumption 4 (Generalization) *The training data on which we build the estimator and the test data on which we make the estimation are drawn from the same distribution, also known as the "no covariate shift" assumption [133].*

S3 Appendix. Major causal-inference methods: When to use which estimator?

G-formula also called conditional mean regression [83], g-computation [59], or Q-model [134]. This approach is directly modeling the outcome, also referred to as the response surface: $\mu_{(a)}(x) = \mathbb{E}(Y \mid A = a, \mathbf{X} = x)$

Using an outcome estimator to learn a model for the response surface $\hat{\mu}$ (eg. a linear model), the ATE estimator is an average over the n samples:

$$\hat{\tau}_G(f) = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}(x_i, 1) - \hat{\mu}(x_i, 0) = \frac{1}{n}\sum_{i=1}^{n} \hat{\mu}_{(1)}(x_i) - \hat{\mu}_{(0)}(x_i) \tag{4}$$

This estimator is unbiased if the model of the conditional response surface $\hat{\mu}_{(a)}$ is well-specified. This approach assumes than $Y(a) = \mu_a(X) + \epsilon_a$ with $\mathbb{E}[\epsilon|X] = 0$. The main drawback is the extrapolation of the learned outcome estimator from samples with similar covariates X but different intervention A.

Propensity Score Matching (PSM) To avoid confounding bias, the ignorability assumption 1) requires to contrast treated and control outcomes only between comparable patients with respect to treatment allocation probabilities. A simple way to do this is to group patients into bins, or subgroups, of similar confounders and contrast the two population's outcomes by matching patients inside these bins [57]. However, the number of confounder bins grows exponentially with the number of variables. [132] proved that matching patients on the individual probabilities to receive treatment –propensity scores– is sufficient to verify ignorability. PSM is a conceptually simple method, but has delicate parameters to tune such as choosing a model for the propensity score, deciding what is the maximum distance between two potential matches (the caliper width), the number of matches by sample, and matching with or without replacement. It also prunes data not meeting the caliper width criteria, and suffers from high estimation variance in highly-dimensional data where extreme propensity weights are common. Finally, the simple bootstrap confidence intervals are not theoretically grounded [135] making PSM more difficult to use for applied practitioners.

Inverse Propensity Weighting (IPW)

A simple alternative to propensity score matching is to weight the outcome by the inverse of the propensity score [58]. It relies on a similar idea as matching but automatically builds a balanced population by reweighting the outcomes with the propensity score model $\hat{e}$ to estimate the ATE:

$$\hat{\tau}_{IPW}(\hat{e}) = \frac{1}{n}\sum_{i=1}^{N} \frac{A_i Y_i}{\hat{e}(X_i)} - \frac{(1 - A_i)Y_i}{(1 - \hat{e}(X_i))} \tag{5}$$

This estimate is unbiased if $\hat{e}$ is well-specified. IPW suffers from high variance if some weights are too close to 0 or 1. In high dimensional cases where poor overlap between treated and control is common, one can clip extreme weights to limit estimation instability.

Doubly Robust Learning, DRL also called Augmented Inverse Probability Weighting (AIPW) [136].

The underlying idea of DRL is to combine the G-formula and IPW estimators to protect against a mis-specification of one of them. It first requires to estimate the two nuisance parameters: a model for the intervention $\hat{e}$ and a model for the outcome f. If one of the two nuisance is unbiased, the following ATE estimator is as well:

$$\widehat{\tau}_{AIPW} = \frac{1}{n}\sum_{i=1}^{n} \left(\hat{\mu}_{(1)}(x_i) - \hat{\mu}_{(0)}(x_i) + a_i \frac{y_i - \hat{\mu}_{(1)}(x_i)}{\hat{e}(x_i)} - (1 - a_i) \frac{y_i - \hat{\mu}_{(0)}(x_i)}{1 - \hat{e}(x_i)} \right)$$

Moreover, despite the need to estimate two models, this estimator is more efficient in the sense that it converges quicker than single model estimators [82]. For this propriety to hold, one need to fit and apply the two nuisance models in a cross-fitting manner. This means that we split the data into K folds. Then for each fold, we fit the nuisance models on the K-1 complementary folds, and predict on the remaining fold.

To recover Conditional Treatment Effects from the AIPW estimator, [137] suggested to regress the Individual Treatment Effect estimates from AIPW on potential sources of heterogeneity X^{cate}: $\hat{tau} = \arg\min_{\tau \in \Theta}(\hat{\tau}_{AIPW}(X) - \tau(X^{cate}))$ for Θ some class of model (eg. linear model).

Double Machine Learning [53] also known as the R-learner [138]. It is based on the R-decomposition, [139], and the modeling of the conditional mean outcome, $m(x) = \mathbb{E}[Y|X = x]$ and the propensity score, $e(x) = \mathbb{E}[A = 1|X = x]$:

$$y_i - m\left(x_i\right) = \left(a_i - e\left(x_i\right)\right)\tau\left(x_i\right) + \varepsilon_i \quad \text{with } \varepsilon_i = y_i - \varepsilon\left[4_i \mid x_i, a_i\right] \tag{6}$$

Note that we can impose that the conditional treatment effect $\tau(x)$ only relies on a subset of the features, x^{cate} on which we want to study treatment heterogeneity.

From this decomposition, we can derive an estimation of the ATE τ, where the right hand-side term is the empirical R-Loss:

$$\hat{\tau}(\cdot) = \text{argmin}_{\tau}\left\{\frac{1}{n}\sum_{i=1}^{n}\left(\left(y_i - m\left(x_i\right)\right) - \left(a_i - e(x_i)\right)\tau\left(x_i^{cate}\right)\right)^2\right\} \tag{7}$$

The full procedure for R-learning is first to fit the nuisances: $\hat{m}$ and $\hat{e}$. Then, minimize the estimated R-loss eq.7, where the oracle nuisances (e, m) have been replaced by their estimated counterparts $(\hat{e}, \hat{m})$. Minimization can be done by regressing the outcome residuals weighted by the treatment residuals. Finally, get the ATE by averaging conditional treatment effect $\tau(x^{cate})$ over the population.

This estimator has also the doubly robust proprieties described for AIPW. it should have less variance than AIPW since it does not use the propensity score in the denominator.

	estimation_method	compute_time	outcome_model	event_aggregation
2	LinearDML	1127.977827	Forests	['first', 'last']
3	backdoor.propensity_score_matching	199.765587	Forests	['first', 'last']
4	backdoor.propensity_score_weighting	86.149872	Forests	['first', 'last']
5	TLearner	284.066786	Forests	['first', 'last']
6	LinearDRLearner	2855.403709	Forests	['first', 'last']
7	LinearDML	49.911035	Regularized LR	['first', 'last']
8	backdoor.propensity_score_matching	127.929910	Regularized LR	['first', 'last']
9	backdoor.propensity_score_weighting	6.407206	Regularized LR	['first', 'last']
10	TLearner	6.843931	Regularized LR	['first', 'last']
11	LinearDRLearner	80.747301	Regularized LR	['first', 'last']

Table 3. *Compute times for the different estimation methods with 50 bootstrap replicates.*

S4 Appendix. Statistical considerations when implementing estimation.

Counterfactual prediction lacks off-the-shelf cross-fitting estimators

Doubly robust methods use cross-fit estimation of the nuisance parameters, which is not available off-the-shelf for IPW and T-Learner estimators. For reproducibility purposes, we did not reimplement internal cross-fitting for treatment or outcome estimators. However, when flexible models such as random forests are used, a fairer comparison between single and double robust methods should use cross-fitting for both. This lack in the scikit-learn API [97] reflects different needs between purely predictive machine learning focused on generalization performances and counterfactual prediction aiming at unbiased inference on the input data.

Good practices for imputation not implemented in EconML

Good practices in machine learning recommend to input distinctly each fold when performing cross-fitting [2]. However, EconML estimators test for missing data at instantiation preventing the use of scikit-learn imputation pipelines. We thus have been forced to transform the full dataset before feeding it to causal estimators. An issue mentioning the problem has been filed, so we can hope that future versions of the package will comply with best practices. [3]

Bootstrap may not yield the most efficient confidence intervals

To ensure a fair comparison between causal estimators, we always used bootstrap estimates for confidence intervals. However, closed form confidence intervals are available for some estimators – see [82] for IPW and AIPW (DRLeaner) variance estimations. These formulas exploit the estimator properties, thus tend to have smaller confidence intervals. On the other hand, they usually do not include the variance of the outcome and treatment estimators, which is naturally dealt with in bootstrapped confidence intervals. Closed form confidence intervals are rarely implemented in any of the packages as Dowhy for the IPW estimator, or in EconML for AIPW.

Bootstrap was particularly costly to run for the EconML doubly robust estimators (AIPW and Double ML), especially when combined with random forest nuisance estimators (from 10 to 47 min depending on the aggregation choice and the estimator). See Table 3 for details.

[2] https://scikit-learn.org/stable/modules/compose.html#combining-estimators
[3] https://github.com/py-why/EconML/issues/664

S5 Appendix. Packages for causal estimation in the python ecosystem. We searched for causal inference packages in the python ecosystem. The focus was on the identification methods. Important features were ease of installation, sklearn estimator support, sklearn pipeline support, doubly robust estimators, confidence interval computation, honest splitting (cross-validation), Targeted Maximum Likelihood Estimation. These criteria are summarized in 4. We finally chose EconML despite lacking `sklearn._BaseImputer` support through the `sklearn.Pipeline` object as well as a TMLE implementation.

The zEpid package is primarily intended for epidemiologists. It is well documented and provides pedagogical tutorials. It does not support sklearn estimators, pipelines and honest splitting.

EconML [96] implements almost all estimators except propensity score methods. Despite focusing on Conditional Average Treatment Effect, it provides all. One downside is the lack of support for scikit-learn pipelines with missing value imputers. This opens the door to information leakage when imputing data before splitting into train/test folds.

Dowhy [140] focuses on graphical models and relies on EconML for most of the causal inference methods (identifications) and estimators. Despite, being interesting for complex inference –such as mediation analysis or instrumental variables–, we considered that it added an unnecessary layer of complexity for our use case where a backdoor criterion is the most standard adjustment methodology.

Causalml implements all methods, but has a lot of package dependencies which makes it hard to install.

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 4. *Selection criteria for causal python packages*

S6 Appendix. Hyper-parameter search for the nuisance models. We followed a two-step procedure to train the nuisance models (eg. $(\hat{e}, \hat{\mu})$ for the AIPW causal estimator), taking inspiration from the computationally cheap procedure from [141, section 3.3]. First, for each nuisance model, we fit a random parameter search with 5-fold cross validation and 10 iterations on the full dataset. Each iteration fit a model with a random combination of parameters in a predefined grid, then evaluate the performance by cross-validation. The best hyper-parameters $\hat{\lambda}^\star$ are selected as the ones reaching the minimal score across all iterations. Then, we feed this parameters to the causal estimator. The single robust estimators (matching, IPW and TLearner) refit the corresponding estimator only once on the full dataset, then estimate the ATE. The doubly-robust estimators use a cross-fitting procedure (K=5) to fit the

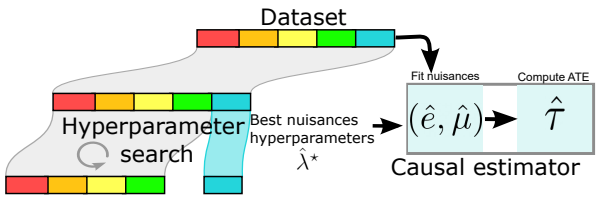


Fig 9. *Hyper-parameter search procedure.*

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 5. *Hyper-parameter grid used during random search optimization.*

nuisances then estimate the ATE. Figure 9 illustrates the procedure and Table 5 details the hyper-parameters grid for the random search.

S7 Appendix. Computing resources.

The whole project was run on a laptop running Ubuntu 22.04.2 LTS with the following hardware: CPU 12th Gen Intel(R) Core(TM) i7-1270P with 16 threads and 15 GB of RAM.

S5 Fig. Selection flowchart.

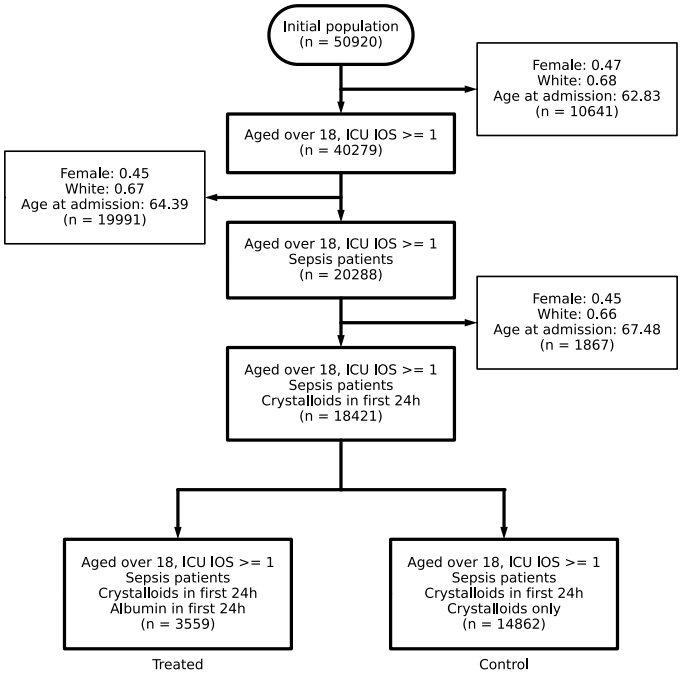


Fig 10. *Selection flowchart on MIMIC-IV for the emulated trial.*

S1 Table. Complete description of the confounders for the main analysis.

	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 6. *Characteristics of the trial population measured on the first 24 hours of ICU stay. Risk scores (AKI, SOFA, SAPSII) and lactates have been summarized as the maximum value during the 24 hour period for each stay. Total cumulative urine output has been computed. Other variables have been aggregated by taking mean during the 24 hour period.*

S6 Fig. Directed Acyclic Graph.

The expert DAG in figure depicts the known causal links between these variables.

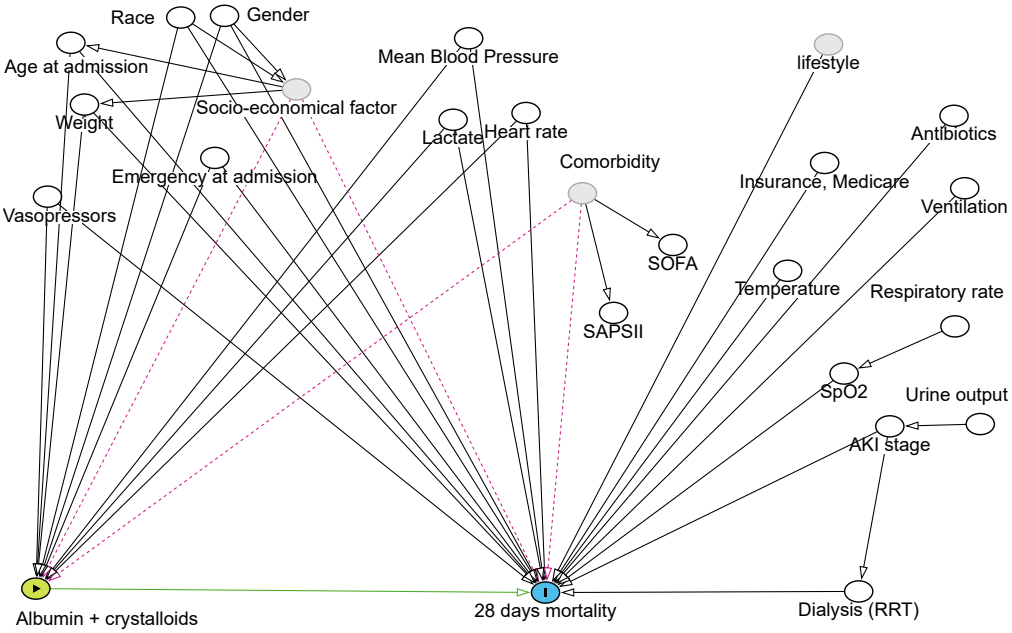


Fig 11. ***Causal graph for the Albumin vs crystalloids emulated trial*** – *The green arrow indicates the effect studied. Black arrows show causal links known to medical expertise. Dotted red arrows highlight confounders not directly observed. For readability, we draw only the most important edges from an expert point of view. All white nodes correspond to variables included in our study.*

S7 Fig. Complete results for the main analysis.

Compared to Figure 3, we also report in Figure 12 the estimates for Causal forest estimators and other choices of feature aggregation (first and last).

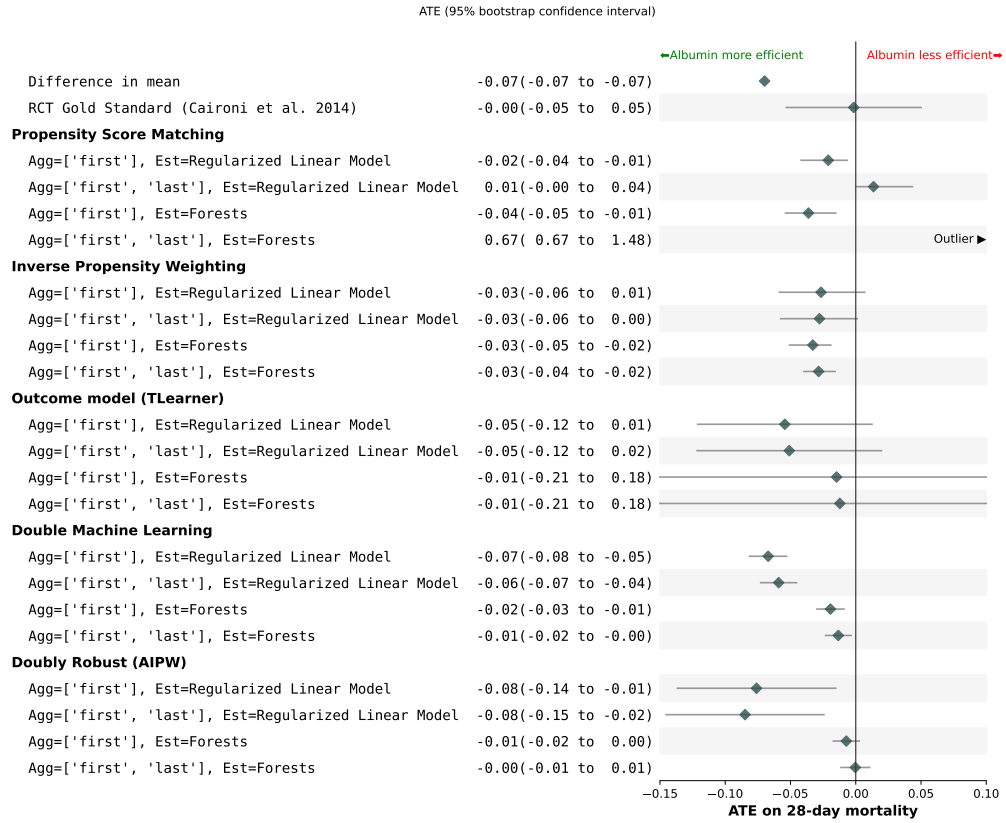


Fig 12. *Full sensitivity analysis: The estimators with forest nuisances point to no effect for almost every causal estimator consistently with the RCT gold standard. Only matching with forest yields an unconvincingly high estimate. Linear nuisance used with doubly robust methods suggest a reduced mortality risk for albumin. The choices of aggregation only marginally modify the results expect for propensity score matching. The green diamonds depict the mean effect and the bar are the 95% confidence intervals obtained by 50 bootstrap repetitions.*

S8 Fig. Complete results for the Immortal time bias.

Compared to Figure 3a, we also report in Figure 13 the estimates for Double Machine Learning, Inverse Propensity Weighting for both Random Forest and Ridge Regression. Feature aggregation was concatenation of first and last for all estimates.

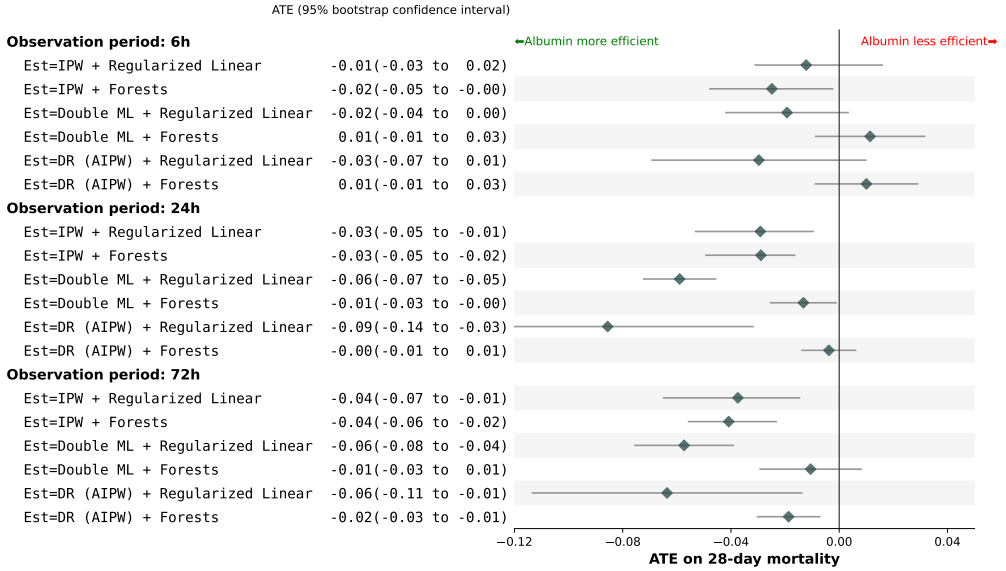


Fig 13. *Sensitivity analysis for immortal time bias: Every choice of estimates show an improvement of the albumin treatment when increasing the observation period, thus increasing the blank period between inclusion and administration of albumin. Aggregation was concatenation of first and last features. The green diamonds depict the mean effect and the bar are the 95% confidence intervals obtained by 50 bootstrap repetitions.*

S8 Appendix. Deviating from expert ignorability – Impact of smaller confounders sets.

We conducted a dedicated vibration analysis on the different choices of confounders. We created three confounder subsets in addition to all confounders (24 variables): all confounders without antibiotics (Glycopeptides, Beta-lactams, Carbapenems, Aminoglycosides), all confounders without any measurement (weight, lactate, heart rate, spo2, mbp, urine output, temperature, AKI stage, SAPSII, respiratory rate, SOFA), only socio-demographics (admission age, female, emergency admission, insurance–medicare, race).

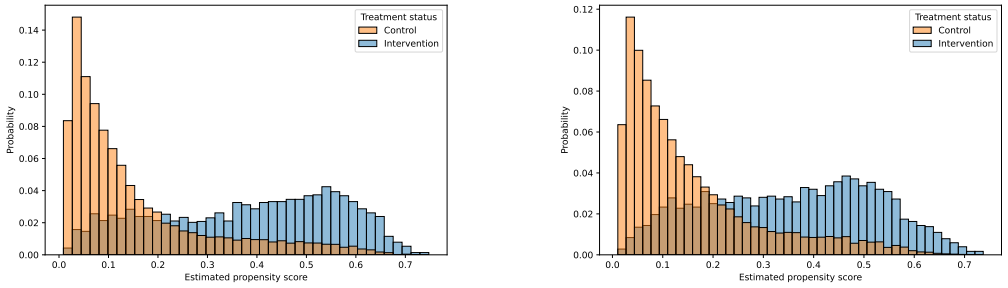
Figure 3b shows that small deviation from the ignorability assumptions is tolerable: for example, removing antibiotics does not impact the estimate. However, the larger the deviation from graph 11, the larger the bias compared to the gold-standard. Adjusting only for socio-demographics features is the closest from an unadjusted risk difference, indicating that we lack important confounders on the patient health state. This stability of the treatment effect estimator once sufficient confounders have been included has already been described and suggested as a confounder selection method [50].

S9 Fig. Vibration analysis for aggregation.

We conducted a dedicated vibration analysis on the different choices of features aggregation, studying the impact on the estimated ATE. We also studied if some choices of aggregation led to substantially poorer overlap.

We assessed overlap with two different methods. As recommended by [58], we did a graphical assessment by plotting the distribution of the estimated. The treatment model hyper-parameters were chosen by random search, then predicted propensity scores were obtained by refitting this estimator with cross-fitting on the full dataset.

As shown in Figure 14, we did not find substantial differences between methods when plotting graphically the distribution of the estimated propensity score.



(a) **(b)**

Fig 14. *Different choices of aggregation yield qualitatively close distributions of the propensity score: Figure 14a)a) shows a concatenation of first, last and median measures whereas Figure 14b)b) shows an aggregation by taking the first measure only. The underlying treatment effect estimator is a random forest.*

We also used normalized total variation (NTV) as a summary statistic of the estimated propensity score to measure the distance between treated and control population [85]. This statistic varies between 0 – perfect overlap – and 1 – no overlap at all. Fig 15 shows no marked differences in overlap as measured by NTV between aggregation choices, comforting us in our expert-driven choice of the aggregation: a concatenation of first and last feature observed before inclusion time.

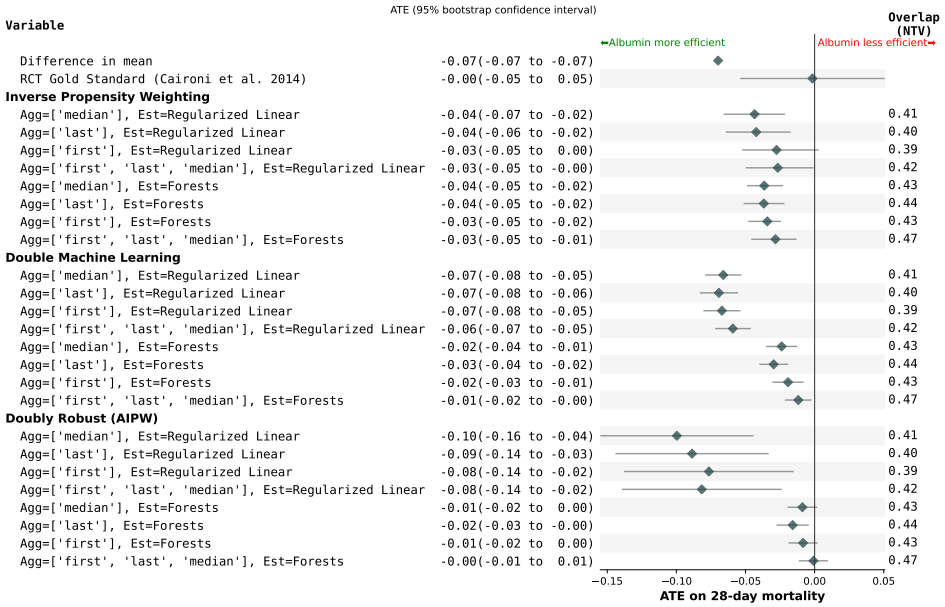


Fig 15. *Vibration analysis dedicated to the aggregation choices. The choices of aggregation only marginally modify the results. When assessed with Normalized Total Variation, the overlap assumption is respected for all our choices of aggregation. The green diamonds depict the mean effect and the bar are the 95% confidence intervals obtained by 50 bootstrap repetitions.*

S8 Appendix. Details on treatment heterogeneity analysis. Detailed estimation procedure

The estimation of heterogeneous effect based on Double Machine Learning adds another step after the computation, regressing the residuals of the outcome nuisance $\tilde{Y} - \mu(X)$ against the residuals of the treatment nuisance $\tilde{A} = A - e(X)$ with the heterogeneity features X_{CATE}. Noting the final CATE model θ, Double ML solves:

$$\arg\min_{\theta} \mathbb{E}_n\left[(\tilde{Y} - \tau(X_C ATE) \cdot \tilde{A})^2\right]$$

Where $\tilde{Y} = Y - \hat{m}(X)$ and $\tilde{A} = A - \hat{e}(X)$

To avoid the over-fitting of this last regression model, we split the dataset of the main analysis into a train set (size=0.8) where the causal estimator and the final model are learned, and a test set (size=0.2) on which we report the predicted Conditional Average Treatment Effects.

Known heterogeneity of treatment for the emulated trial

[94] observed statistical differences in the post-hoc subgroup analysis between patient with and without septic shock at inclusion. They found increasing treatment effect measured as relative risk for patients with septic shock (RR=0.87; 95% CI, 0.77 to 0.99 vs 1.13;95% CI, 0.92 to 1.39).

[142] conducted a post-hoc subgroup analysis of patients with or without brain injury –defined as Glasgow Coma Scale between 3 to 8–. The initial population was patients with traumatic brain injury (defined as history or evidence on A CT scan of head trauma, and a GCS score <= 13). They found higher mortality rate at 24 months in the albumin group for patients with severe head injuries.

[95] conducted a subgroup analysis on age (<60 vs >60), septic shock and sex. They conclude for increasing treatment effect measured as Restricted Mean Survival Time for Sepsis vs septic shock (3.47 vs. 2.58), for age >=60 (3.75 vs 2.44), for Male (3.4 vs 2.69). None of these differences were statistically significant.

Vibration analysis

The choice of the final model for the CATE estimation should also be informed by statistical and clinical rationals. Figure 16 shows the distribution of the individual effects of a final random forest estimator, yielding CATE estimates that are not consistent with the main ATE analysis. Figure 17 shows that the choice of this final model imposes an inductive bias on the form of the heterogeneity and different sources of noise depending of the nature of the model. A random forest is noisier than a linear model. Figure 17 shows the difference of modelization on the subpopulation of non-white male patients without septic shock. One can see that the decreasing linear trend is reflected by the random forest model only for patients aged between 55 and 80.

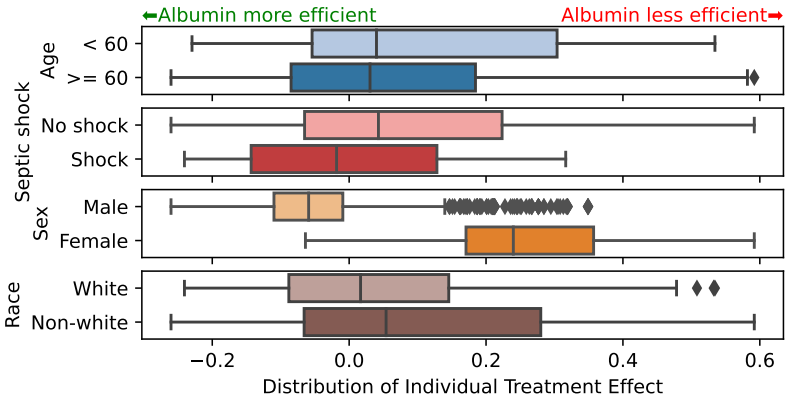


Fig 16. *Distribution of Conditional Average Treatment effects on sex, age, race and pre-treatment septic shock estimated with a final forest estimator. The CATE are positive for each subgroups, which is not consistent with the null treatment effect obtained in the main analysis. The boxes contain between 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.*

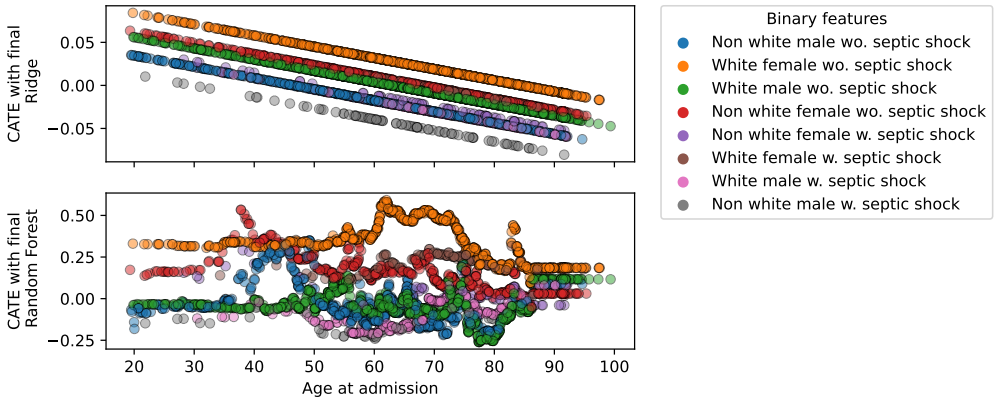


Fig 17. *Distribution of Conditional Average Treatment effects on sex, age, race and pre-treatment septic shock plotted for different ages. On the top the final estimator is a linear model; on the bottom, it is a random forest. The forest-based CATE displays more noisy trends than the linear-based CATE. This suggest that the flexibility of the random forest might be underfitting the data.*

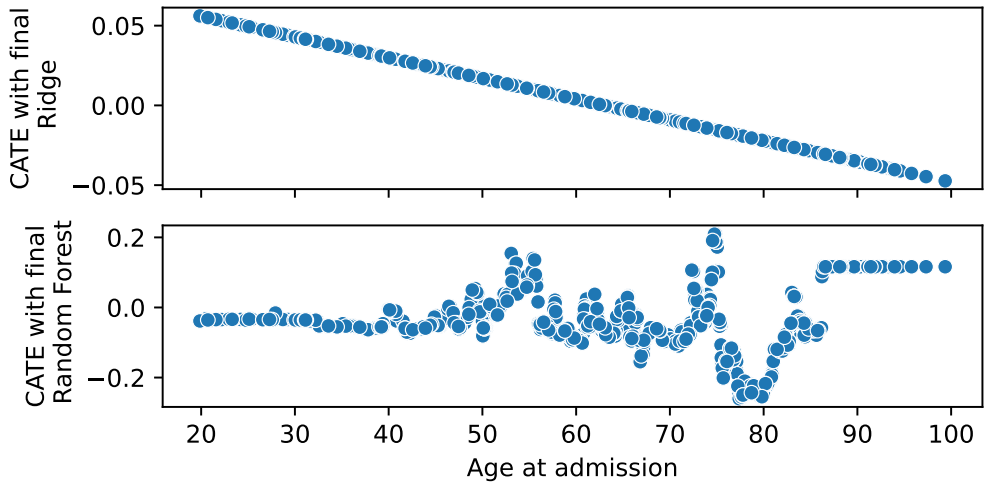


Fig 18. *Figure 17 on the subpopulation of white male patients without septic shock. Contrary to the ridge regression (on top) inducing a nicely interpretable trend, using random forests as the final estimator failed to recover CATE on ages: the predicted estimates do not exhibit any trend and display inconsistently large effect sizes, suggesting data underfitting.*

S9 Appendix. Risk assessment tools for prevention.

An example of nationally deployed cardiovascular prevention tool: QRISK

The QRISK is a risk score widely used in the UK to assess 10-year cardio-vascular risk [24]. It is well calibrated and have satisfactory discriminative performances. As such, it is recommended by the national regulatory agency to engage the patient into various preventive interventions such as physical activity, cardio-protective diet, statins initiation [25]. The NICE argues in favor of systematic screening of cardio-vascular risks to increase the number of people receiving statins [25].

The QRISK seems to mostly inform the initiation of statins for patients exceeding 10% 10-year CVD risk (as measured with QRISK) following the 2014 NICE guidelines [26].

They base their evidence on a cost-effectiveness analysis choosing the 10% threshold to optimize quality-adjusted life year at the population level based on the cost of life-long statin treatments [27]. This study assumes constant statins effect among different CVD risk groups: *It is assumed that the risk ratios given for treatment with each class of statins are constant regardless of the baseline CV risk –that is, someone with low CV risk will receive the same proportional reduction in that risk as would someone with a high CV risk. This is unproven, but is consistent with the results of meta-analysis carried out by the Cholesterol Treatment Trialists, which found effectiveness to be broadly similar for those at different risk levels.* This hypothesis is supported by the meta-analysis of at least for the important risk factors of sex and age [31]. It is also assumed that these risk ratios are constant regardless of baseline LDL-cholesterol levels. More recent work showed that risk ratio stratified by patient baseline CVD risk levels were not significant, raising doubts about the overall efficacy of statins for primary prevention [32].

The updated guidelines from the US Preventive Services Task Force found respectively moderate and small net benefice of statins for patients aged 40 to 75 with above 10% CVD risk and from 7.5 to 10% without history of CVD [33]. They conclude that the evidence is insufficient to determine the balance of benefits and harms of statin use for the primary prevention of CVD events and mortality in adults 76 years or older with no history of CVD.

Critics on the efficiency of systematic screening points to treatment heterogeneity

On the other side, the WHO systematic review on systematic screening for cardio-vascular risks from 2019 [34] is categorical on the poor effectiveness of systematic cardiovascular risk screening, consistently with the Cochrane review on poor effectiveness [35]. For the WHO, systematic screening leads to over-diagnosis and over-treatment. They mention in their argumentation no positive effect of statins and even adverse effect for mild hypertension [143]. They also mention critics against poor uptake of the systematic screening in UK or Albany leading to documented social bias in the screened population: for example, in the UK, treatment heterogeneity at a given risk level [36] or documentation of social inequities of health checks [37].

- The first critic shows that the augmentation of statin prescriptions (up to 400% increase for high risk patients), has been associated with heterogeneity in statin prescriptions for given risk score levels [36]. Low risk patients have been over-prescribed and high risk patient under-prescribed, resulting in questionable overall improvement. In this case, the constant effect hypothesis of statin proposed by the NICE cost-effectiveness study is in doubt. Having access to sources of heterogeneity would help to optimize treatment allocation. The risk score might capture interesting heterogeneity directions, but we lack data from RCTs to draw such conclusions.

- The second critic points to a treatment allocation bias for risk assessment. It seems that mostly socially advantaged patients are receiving risk assessment and thus the possibility to be treated with statins. This critic also points to treatment heterogeneity. Suppose that every patient benefits the same from statins, then even a small part of the population –such as the socially advantaged at more than 10% risk– would benefit from it and make the intervention effective at the population level.

NICE evidence for statin efficacy

In the NICE analysis, three trials weight for more than 70% of the effect. All focus on high risk patients and secondary prevention. It is not clear if those effects are used for the cost-effectiveness.

- LIPID [28] (32% of the effect): 9,014 patients with a history of myocardial infraction or hospitalization ie. secondary prevention.
- PROSPER [29] (22% of the effect): an elderly cohort of 5,804 men and women with, or at high risk of developing, cardiovascular disease and stroke.
- ALLHAT-LLT [30] (23% of the effect): 10,355 patients age more than 55 years and stage 1 or 2 hypertension with at least 1 additional CHD risk factor.