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] **1** Inria, Soda, Saclay, France

2 Mission Data, Haute Autorité de Santé, Saint-Denis, France
3 Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, MA 02139
4 Medical University Clinic, Division of Endocrinology, Diabetes & Metabolism, Kantonsspital Aarau, Aarau, Switzerland
5 Agence Régionale de Santé Ile-de-France, France
6 Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, MA 02215
7 Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115

* Corresponding author: m.doutreligne@has-sante.fr

Abstract

Causal inference enables machine learning methods to estimate treatment effects of medical interventions from electronic health records (EHRs). The prevalence of such observational data and the difficulty for randomized trial to cover all population/treatment relationships make these methods increasingly attractive for the study of causal effects.

Here we explain and study common methodological pitfalls and potential solutions. We illustrate our framework for causal inference estimating the effect of albumin on mortality in sepsis using an Intensive Care database (MIMIC-IV) and comparing credible analyses variations to results from randomized controlled trials (RCT) as gold-standard.

The first step is study design, using the target trial concept and the PICOT framework: Population (patients with sepsis), Intervention (combination of crystalloids and albumin for fluid resuscitation), Control (crystalloids only), Outcome (28-day mortality), Time (intervention start within 24h of admission). We show that too large treatment-initiation time induces immortal time bias. The second step is selection of the confounding variables based on expert knowledge. Increasingly adding confounders enables to recover the RCT results from observational data (0%, 95% CI -1% to 1%, and 0%, 95% -5% to 5%, respectively). As the third step, we assess the influence of multiple models with varying assumptions, showing that a doubly robust estimator (AIPW) with random forests proved to be the most reliable estimator. Results show that these steps are all important for valid causal estimates. A valid causal model can then be used to individualize decision making: subgroup analyses showed that treatment efficacy of albumin was better for patients >60 years old, males, and patients with septic shock.

Without causal thinking, machine learning is not enough for optimal clinical decision making for each patient. Our step-by-step analytic framework helps the analyst avoid the many pitfalls of applying machine learning to EHR data and build models that avoid shortcuts and extract the best decision-making evidence.

Author summary

Rich routine-care data, as EHR or claims, is useful to individualize decision making using machine learning; but guiding interventions requires causal inference. Unlike with an RCT, interventions in routine data do not easily enable an apple-to-apple measure of the effect of an intervention, leading to many analytical pitfalls, particularly in time-varying data. We study these in a tutorial spirit, making the code and data openly available. We give 5 analytical steps for data-driven individualized interventions: Step 1) Study design, where common pitfalls are selection bias, with information unequally collected across treatment and control patients, and immortal time bias, where the inclusion-defining event interacts with the intervention time. Step 2) Identification of the causal assumptions and categorization of confounders. Step 3) Estimation of the causal effect of interest by correct aggregation of confounders and selection of an appropriate statistical model. Step 4) Assessing the analysis' robustness to assumptions, and finally Step 5) Individualizing treatment decision, by exploring treatment heterogeneity, eg across subgroups. Studying choice of fluid resuscitation in sepsis, we show that common mistakes in steps 1, 2, and 3 equally compromise causal validity.

Introduction: data-driven decisions require causal inference

Informing a care option extends beyond merely predicting the occurrence of an event; it involves estimating the effect of the corresponding treatment effects. Routine-care data comes naturally to mind to guide routine decisions, but they require care to estimate treatment effects as they are observational, unlike Randomized controlled trials (RCTs). This context calls for causal inference statistical frameworks. But merely applying these tools to the data does suffice to ensure the validity of the inferences; numerous considerations must be carefully addressed.

Individualized Medicine and Machine Learning Challenges Machine learning plays a pivotal role in individualized medicine [1–5]. It demonstrated superior performance over traditional rule-based clinical scores in predicting a patient's readmission risk, mortality, or future comorbidities using Electronic Health Records (EHRs) [1–5]. However, mounting evidence suggests that machine-learning models can inadvertently perpetuate and exacerbate biases present in the data [6], including gender or racial biases [7, 8], and the marginalization of under-served populations [9]. These biases are typically encoded by capturing shortcuts—stereotypical or distorted features in the data [10–12]. For instance, numerous machine learning algorithms rely on post-treatment information [13–16], exemplified by a diagnostic model for skin cancer that depends on surgical marks [11]. For Intensive Care Unit data, focus of our study, such information markedly improves mortality prediction (Figure S1 5), but cannot inform decisions.

The Importance of Causal Reasoning in Data-Driven Decision-Making [17] While conventional machine learning relies on retrospective to generate predictions of future effects [18], truly informing decision-making needs a comparison of potential outcomes with and without the intervention. This involves estimating a causal effect, mirroring the methodology employed in RCTs [17]. However, RCTs encounter challenges such as selection biases [19, 20], difficulties in recruiting diverse populations, and limited sample sizes for exploring treatment heterogeneity across subgroups. Routinely collected data presents a unique opportunity to assess real-life benefit-risk trade-offs associated

with a decision [21], with reduced sampling bias and sufficient data to capture heterogeneity [22]. Nevertheless, estimating causal effects from such data is challenging due to the confounding of the intervention by indication. Therefore, dedicated statistical techniques are imperative to emulate a "target trial" [23] from observational data.

Critics on the efficiency of systematic screening and treatment heterogeneity Conversely, the World Health Organization (WHO) published a systematic review on screening for CVD risk in 2019 [24] asserting the limited effectiveness of such screenings, aligning with a Cochrane review that also reported poor effectiveness [25] resulting in over-treatment. The surge in statin prescriptions has been linked to a divergence in statin prescriptions at specified risk score levels [26]. This leads to over-prescription for low-risk patients and under-prescription for high-risk patients, casting doubt on the overall efficacy as the constancy assumption of the statin treatment effect might be violated. Accounting for sources of heterogeneity in patient profiles would be instrumental in optimizing treatment allocation.

Moreover, the WHO report critiques suboptimal adoption leading to social biases in the screened populations. In the UK, there are instances of treatment heterogeneity at a given risk level [26] and documented evidence of social inequities in health checks [27]. Evidence suggests that predominantly socially advantaged patients undergo screening, consequently being offered statin treatment. This critique further emphasizes the presence of treatment heterogeneity.

Multiple Perspectives on Evidence-based Decision Making Across different fields, existing literature has emphasized different challenges associated with estimating treatment effects using observational data. While epidemiologic studies underscore the importance of the target trial approach [28–32], there emphasis primarily lies on biases that arise from temporal effects [23, 33–37] or confounding variables [38–40], with relatively less attention to issues arising from estimator selection. Recent replications of RCTs using observational data did not explore the impact of modern machine learning methods on the robustness of the results [31, 41].

In contrast, machine learning and causal inference literature predominantly studies estimators [42–46] : propensity score matching [47], inverse probability weighting [48], outcome models [49], doubly robust methods, [43] or deep learning based models [50]. This literature may be opaque for some due to intricate mathematical details and unverifiable assumptions. Guidelines seldom address time-related biases, or covariate aggregation which frequently emerge in datasets with temporal dependencies [33, 35]. Recently, the machine learning community shifted its focus from EHR data to simulated data, which may not capture the complexities of real-world data [51–54].

In this work, we bring together epidemiological concepts and principles from statistical and machine learning literature. We adopt an empirical perspective to answer practical needs of applied researchers. A study of choices spread out across the analysis –study design, consideration of confounders, and selection of estimators (refer to Section Step-by-step framework for robust decision-making from EHR data)– highlights their equal importance in ensuring the validity of results. To illustrate and compare biases, we investigate the impact of albumin on sepsis mortality using data from a publicly available intensive care database, MIMIC-IV [55] (section Application: evidence from MIMIC-IV on which resuscitation fluid to use).

The primary focus of the main section is on accessibility, with technical details expanded in the appendices.

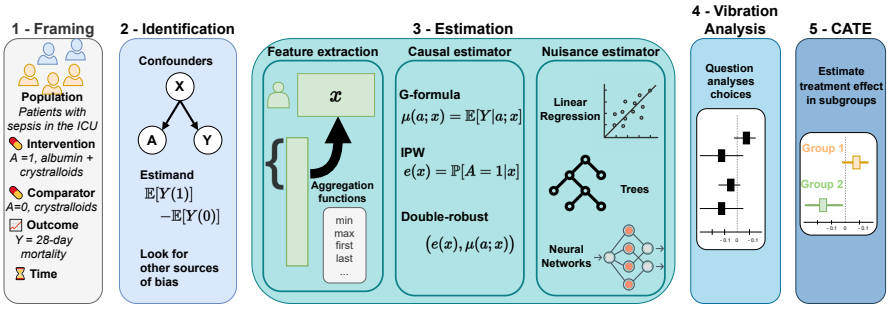


Fig 1. ***Step-by-step analytic framework*** – *The complete inference pipeline confronts the analyst with many choices, some guided by domain knowledge, others by data insights. Making those choices explicit is necessary to ensure robustness and reproducibility.*

Step-by-step framework for robust decision-making from EHR data

Whether or not using machine learning, many pitfalls threaten an analysis' value for decision-making. To avoid these pitfalls, we outline a simple step-by-step analytic framework illustrated in Figure 1 for retrospective case-control studies. We frame the medical question as a target trial [56] to match the design to an RCT giving the gold standard average effect. Then we probe for heterogeneity –predictions on sub-groups– going beyond what RCTs can achieve.

Step 1: study design – Frame the question to avoid biases

Grounding decisions on evidence needs well-framed questions, defined by their PICO(T) components. Population, Intervention, Control, and Outcome [57, 58], and in case of EHRs or claims data an additional time component, are necessary to concord with a (hypothetical) target randomized clinical trial [41, 59] – Table 1. A selection flowchart such as in S5 Fig makes the inclusion and exclusion for the PICOT choices explicit.

Without care in defining these PICO(T) components, non-causal associations between treatment and outcomes can easily be introduced into an analysis [60]. The time-varying nature of EHR calls for checking systematically of the Population and Time components by addressing two commonly encountered types of bias.

Selection Bias: In EHRs, outcomes and treatments are often not directly available and need to be inferred from indirect events. These signals could be missing not-at random, sometimes correlated with the treatment allocation [61]. For example, billing codes can be strongly associated with case-severity and cost. Consider comparing the effectiveness of fluid resuscitation with albumin to crystalloids. As albumin is more costly, this treatment is more likely to have a sepsis billing code. On the contrary, for patients treated with crystalloids, only the most severe cases will have a billing code. Naively comparing patients would overestimate the effect of albumin.

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

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

Immortal time bias: Improper alignment of the inclusion defining event and the intervention time is a major source of bias in time-varying data [23, 33, 36]. Immortal time bias (illustrated in Appendix 6) occurs when the follow-up period, i.e. cohort entry, starts before the intervention, e.g. prescription for a second-line treatment. In this case, the treated group will be biased towards patients still alive at the time of assignment and thus overestimating the effect size. Other frequent temporal biases are lead time bias [34, 35] or right censorship [23], and attrition bias [37]. Good practices include explicitly stating the cohort inclusion event [62, Chapter 10:Defining Cohorts] and defining an appropriate grace period between starting time and the intervention assignment [23]. At this step, a population timeline can help.

Step 2: identification – List necessary information to answer the causal question

The identification step builds a causal model to answer the research question. Indeed, the analysis must compensate for differences between treated and non-treated that are not due to the intervention ([63, chapter 1], [30, chapter 1]).

Causal Assumptions Valid causal inference requires assumptions [64] –detailed in S1 Appendix. The analyst should thus review the plausibility of the following: 1) Unconfoundedness: after adjusting for the confounders as ascertained by domain expert insight, treatment allocation should be random; 2) Overlap –also called positivity– the distribution of confounding variables overlaps between the treated and controls –this is the only assumption testable from data [48]–; 3) No interference between units and consistency in the treatment, a reasonable assumption in most clinical questions.

Categorizing covariates Potential predictors –covariates– should be categorized depending on their causal relations with the intervention and the outcome (illustrated in S4 Fig): *confounders* are common causes of the intervention and the outcome; *colliders* are caused by both the intervention and the outcome; *instrumental variables* are a cause of the intervention but not the outcome, *mediators* are caused by the intervention and is a cause of the outcome. Finally, *effect modifiers* interact with the treatment, and thus modulate the treatment effect in subpopulations [65].

To capture a valid causal effect, the analysis should only include confounders and possible treatment-effect modifiers to study the resulting heterogeneity. Regressing the outcome on instrumental and post-treatment variables (colliders and mediators) will lead to biased causal estimates [39]. Drawing causal Directed Acyclic Graphs (DAGs) [38], *eg* with a webtool such as DAGitty [66], helps capturing the relevant variables and defining a suitable estimand or effect measure. Unconfoundedness –inclusion of all confounders in the analysis– is a strong hypothesis that can be hard to obtain in real applications. In these cases, sensitivity analyses for omited variable bias allow to test the robustness of the results to missing confounders [67], proximal inference can be used to leverage proxy of unobserved confounders [68], and the presence of natural experiment might identify the desired causal effect without unconfoundedness [69, Chapter 5, 9].

The *estimand* is the final causal quantity estimated from the data. Depending on the question, different estimands are better suited to contrast the two potential outcomes E[Y(1)] and E[Y(0)] [70, 71]. For continuous outcomes, risk difference is a natural estimand, while for binary outcomes (e.g. events) the choice of estimand depends on the scale. Whereas the risk difference is very informative at the population level, e.g. for medico-economic decision-making, the risk ratio and the hazard ratio are more informative at the level of sub-groups or individuals [71].

Causal estimators A given estimand can be estimated through different methods. One can model the outcome with regression models also known as G-formula, [49] and use it as a predictive counterfactual model for all possible treatments for a given patient. Alternatively, one can model the propensity of being treated use it for matching or Inverse Propensity Weighting (IPW) [48]. Finally, doubly robust methods model both the outcome and the treatment, benefiting from the convergence of both models [69]. There is a variety of doubly robust models, reviewed in S2 Appendix.

Step 3: Statistical estimation – Compute the causal effect of interest

Confounder aggregation Confounders captured via measures collected over multiple time points must be aggregated at the patient level. Simple forms of aggregation include taking the first or last value before a time point, or an aggregate such as mean or median over time. More elaborate choices may rely on hourly aggregations providing more detailed information on the disease course such as vital signs. They may reduce confounding bias between rapidly deteriorating and stable patients but also increase the number of confounders making estimation more challenging [72]. The increase of variance occurs either in arbitrarily small propensity scores for treatment models or in hazardous extrapolation from one group to another for outcome model. If multiple choices appear reasonable, one should compare them in a vibration analysis (see Step 4: Vibration analysis – Assess the robustness of the hypotheses).

Beyond tabular data, unstructured clinical text may capture confounding or prognostic information [73, 74] which can be added in the causal model [32]. However, high-dimensional confounder space such as text may break the positivity assumption just as hourly aggregation choices for measurements.

Missing covariate values might also be a source of confounding. Some statistical estimators (such as forests) can directly incorporate them as supplementary covariates. Others, such as linear models, require imputations. S3 Appendix details general sanity check for imputation strategies when using statistical estimators.

Statistical estimation models of outcome and treatment The causal estimators use models of the outcome or the treatment –called nuisances. There is currently no clear best practice to choose the corresponding statistical model [52, 75]. The trade-off lies between simple models risking misspecification of the nuisance parameters versus flexible models risking to overfit the data at small sample sizes. Stacking models of different complexity as in a super-learner is a good solution to navigate the trade-off [76, 77].

Step 4: Vibration analysis – Assess the robustness of the hypotheses

Some choices in the pipeline may not be clear cut. Several options should then be explored, to derive conceptual error bars going beyond a single statistical model. When quantifying the bias from unobserved confounders, this process is sometimes called sensitivity analysis [78–80]. Following [81], we use the term vibration analysis to describe the sensitivity of the results to all analytic choices.

Step 5: Treatment heterogeneity – Compute treatment effects on subpopulations

Once the causal design and corresponding estimators are established, they can be used to explore the variation of treatment effects among subgroups. A causally-grounded model can be used to predict the effect of the treatment from all the covariates –confounders and effect modifiers– the *Conditional Average Treatment Effect* (CATE) [82]. Practically, CATEs can be estimated by regressing an individual's predictions given by the causal estimator against the sources of heterogeneity (details in S7 Appendix).

Application: evidence from MIMIC-IV on which resuscitation fluid to use

We now use the above framework to extract evidence-based decision rules for resuscitation. Ensuring optimal organ perfusion in patients with septic shock requires resuscitation by reestablishing circulatory volume with intravenous fluids. While crystalloids are readily available, inexpensive and safe, a large fraction of the administered volume is not retained in the vasculature. Colloids offer the theoretical benefit of retaining more volume, but might be more costly and have adverse effects [83]. Meta-analyses from multiple pivotal RCTs found no effect of adding albumin to crystalloids [84, 85] on 28-day and 90-day mortality. Given this previous evidence, we thus expect no average effect of albumin on mortality in sepsis patients. However, studies –RCT [86] and observational [87]– have found that septic-shock patients do benefit from albumin.

Emulated trial: Effect of albumin in combination with crystalloids compared to crystalloids alone on 28-day mortality in patients with sepsis Multiple published RCTs can validate the analysis pipeline before investigating sub-population effects for individualized decisions. Using MIMIC-IV [55], we compare the magnitude of biases introduced by reasonable choices in the different analytical steps recalled in Figure 2.

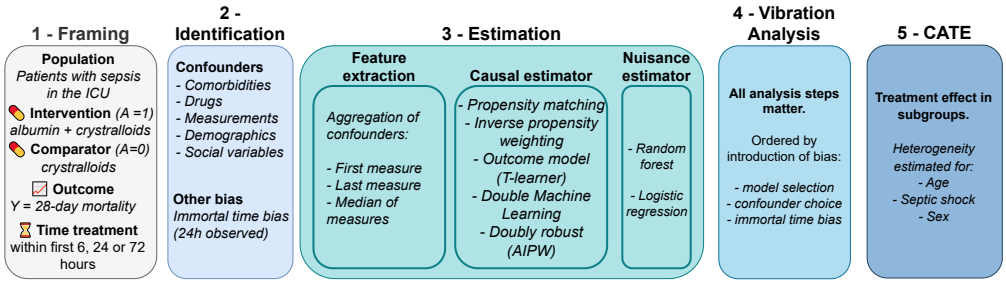


Fig 2. ***Application of the step-by-step framework on which resuscitation fluid to use.***

Study design: effect of crystalloids on mortality in sepsis

- Population: Patients with sepsis in an ICU stay according to the sepsis-3 definition. Other inclusion criteria: sufficient follow-up of at least 24 hours, and age over 18 years. S5 Fig details the selection flowchart and S1 Table the population characteristics.

- Intervention: Treatment with a combination of crystalloids and albumin during the first 24 hours of an ICU stay.
- Control: Treatment with crystalloids only in the first 24 hours of an ICU stay.
- Outcome: 28-day mortality.
- Time: Follow-up begins after the first administration of crystalloids. Thus, we potentially introduce a small immortal time bias by allowing a time gap between follow-up and the start of the albumin treatment –see the full timeline in S3 Fig. Because we are only considering the first 24 hours of an ICU stay, we hypothesize that this gap is insufficient to affect our results. We test this hypothesis in the vibration analysis step.

In MIMIC-IV, these inclusion criteria yield 18,121 patients of which 3,559 were treated with a combination of crystalloids and albumin. While glycopeptide antibiotic therapy was similar between both groups (51.8% crystalloid vs 51.5% crystalloids + albumin), aminoglycosides, carbapenems, and beta-lactams were more frequent in the crystalloid only group (2.0% vs. 0.7%, 4.3% vs. 2.6%, and 35.5% vs. 13.8%, respectively). The crystalloid only group was more frequently admitted as an emergency (57.3% vs. 30.7%). Vasopressors (80.2% vs 41.7%) and ventilation (96.8% vs 87.0%) were more prevalent in the treated populations, underlying the overall higher severity of patients receiving albumin (mean SOFA at admission 6.9 vs. 5.7). Table 2 details patient characteristics.

	Missing	Overall	Cristalloids only	Cristalloids + Albumin	P-Value
n		18421	14862	3559	
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)	
admission_age, mean (SD)	0	66.3 (16.2)	66.1 (16.8)	67.3 (13.1)	<0.001
SOFA, mean (SD)	0	6.0 (3.5)	5.7 (3.4)	6.9 (3.6)	<0.001
lactate, mean (SD)	4616	3.0 (2.5)	2.8 (2.4)	3.7 (2.6)	<0.001

Table 2. *Characteristics of the trial population measured on the first 24 hours of ICU stay. Appendix 6 describes all confounders used in the analysis.*

Identification: listing confounders

For confounders selection we use a causal DAG shown in Figure S6 Fig. Gray confounders are not controlled for since they are not available in the data. However, resulting confounding biases are captured by proxies such as comorbidity scores (SOFA or SAPS II) or other variables (eg. race, gender, age, weight). S1 Table details confounders summary statistics for treated and controls.

Causal estimators: We implemented multiple estimation strategies, including Inverse Propensity Weighting (IPW), outcome modeling (G-formula) with T-Learner, Augmented Inverse Propensity Weighting (AIPW) and Double Machine Learning (DML). We used the python packages dowhy [45] for IPW implementation and EconML [88] for all other estimation strategies. Confidence intervals were estimated by bootstrap (50 repetitions). S2 Appendix and S4 Appendix detail the estimators and the available Python implementations. S3 Appendix details statistical considerations that we identify as important but missing in these packages, namely lack of cross fitting estimators, bad practices for imputation, lack of closed form confidence intervals.

Estimation

Confounder aggregation: We tested multiple aggregations such as the last value before the start of the follow-up period, the first observed value, and both the first and last values as separated features. Missing values were median imputed for numerical features, categorical variables were one-hot encoded (thus discarded missing values).

Outcome and treatment estimators: To model the outcome and treatment, we used two common but different estimators: random forests and ridge logistic regression implemented with scikit-learn [89]. We chose the hyperparameters with a random search procedure (S5 Appendix). While logistic regression handles predictors in a linear fashion, random forests bring the benefit of modeling non-linear relations.

Vibration analysis: Comparing sources of systematic errors

Study design – Illustration of immortal time bias: To illustrate the risk of immortal-time bias, we vary the eligibility period of treatment or control in a shorter or longer time window than 24 hours. As explained in section , a longer eligibility period means that patients are more likely to be treated if they survived up to the intervention and hence the study is biased to overestimate the beneficial effect of the intervention. Figure 3a shows that longer eligibility periods lead to albumin being markedly more efficient (detailed results with causal forest and other choices of aggregation in S8 Fig).

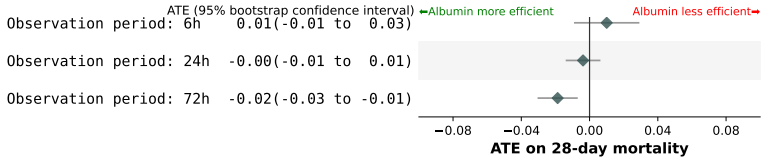
Confounder choice We consider other choice of confounding variables (S6 Appendix). Figure 3b shows that a less thorough choice, neglecting the administrated drugs, makes little to no difference. Major errors, such as omitting the biological measurements or using only socio-demographical variables, lead to sizeable bias. This is consistent with the literature enhancing the importance of a clinically valid DAG [38].

Estimation choices – Confounder aggregation, causal and nuisance estimators: Figure 3c shows varying confidence intervals (CI) depending on the method. Doubly-robust methods provide the narrowest CIs, whereas the outcome-regression methods have the largest CI. The estimates of the forest models are closer to the consensus across prior studies (no effect) than the logistic regression indicating a better fit of non-linear relationships. We only report the first and last pre-treatment feature aggregation strategy, since detailed analysis showed little differences for other aggregations (S7 Fig for complete results, and S9 Fig for a detailed study on aggregation choices). Both methodological studies [90] and consistency with published RCTs suggest to prefer doubly-robust approaches.

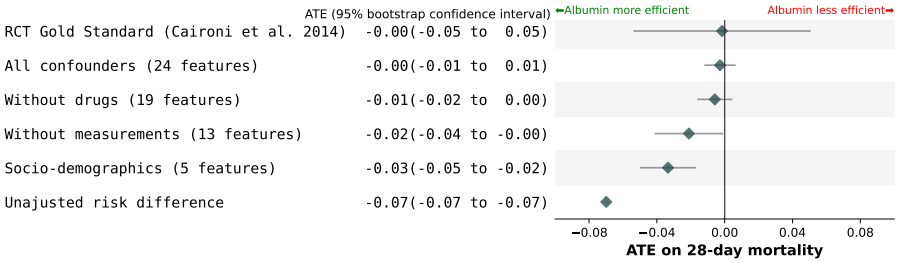
Treatment heterogeneity: Which treatment for a sub-population?

With adequate choice of study design, confounding variables and causal estimator, the average treatment effect matches well published findings: Pooling evidence from high-quality RCTs, no effect of albumin in severe sepsis was demonstrated for both 28-day mortality (odds ratio (OR) 0.93, 95% CI 0.80-1.08) and 90-day mortality (OR 0.88, 95% CI 0.761.01) [84]. Having validated the analytical pipeline, we can use it to inform decision-making. We explore heterogeneity along four binary patient characteristics, displayed in Figure 4. We find that albumin is beneficial with patient with septic shock consistent with one RCT [86]. It is also beneficial for older patients (age >=60) and males. S7 Appendix details the heterogeneity analysis.

(a) ***Framing – Immortal Time Bias***



(b) ***Identification – confounders choice***



(c) ***Model selection***

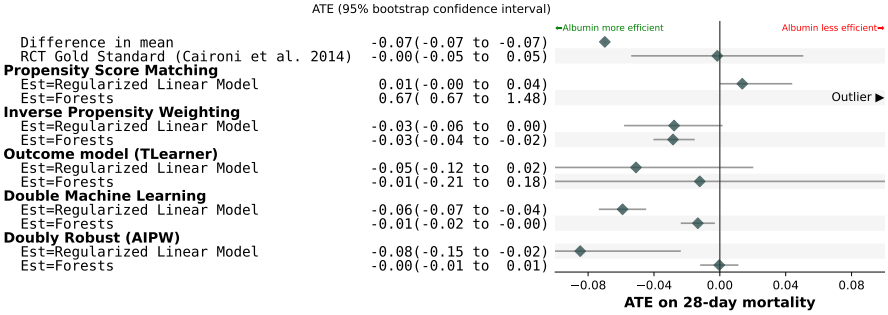


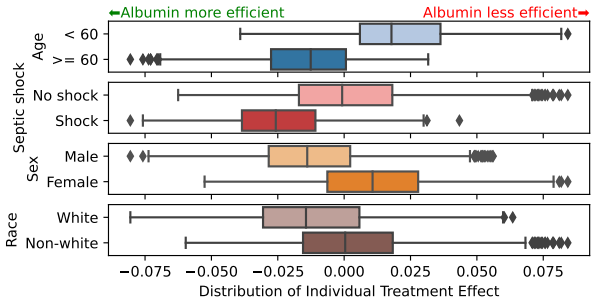
Fig 3. ***The effect of choices on the three analytical steps*** *– All three analytical steps are equally important for the validity of the analysis. 3a) Framing step: Poor framing introduces time bias: A longer observation period (72h) artificially favors the efficacy of Albumin. 3b) Identification step: Choosing less informed confounders set introduces increasing bias in the results. 3c) Model selection step: Different estimators give different results. Score matching yields unconvincingly high estimates, inconsistent with the published RCT. With other causal approaches, using linear estimators for nuisances suggest a reduced mortality risk for albumin, while using forests for nuisance models points to no effect, which is consistent with the RCT gold standard.*
The diamonds depict the mean effect and the bar are the 95% confidence intervals obtained respectively by 30, 30 and 50 bootstrap repetitions. For framing and identification, the estimator is a doubly robust learner (AIPW) with random forests for nuisances. Features are aggregated by taking the first and last measurements for all experiments.

Discussion and conclusion

Valid decision-making evidence from EHR data requires a clear causal framework. Indeed, machine-learning algorithms have often extracted non-causal associations between the intervention and the outcome, improper for decision-making [11, 13, 14]. Machine learning studies in medicine often rely on an implicit causal thinking, via a good understanding of the clinical settings. A clear framework helps making sure nothing falls through the cracks.

We have separated three steps important for causal validity: the choice of study design, confounders, and estimators. Regarding study design, major caveats arise from the time component, where a poor choice of inclusion time easily brings in significant

Fig 4. ***Subgroup distributions of Individual Treatment effects****: better treatment efficacy for patients older than 60 years, septic shock, and to a lower extent males. The final estimator is ridge regression. The boxes contain the* 25^{th} *and* 75^{th} *percentiles of the CATE distributions with the median indicated by the vertical line. The whiskers extend to 1.5 times the inter-quartile range of the distribution.*



bias. Regarding choice of prediction variables, forgetting some variables that explains both the treatment allocation and the outcome leads to confounding bias, that however remains small when these variables capture weak links. Regarding choice of causal estimators, preferring flexible models such as random forests reduces the bias, in particular for doubly-robust estimators. We have shown that all these three steps are equally important: paying no attention to one of them leads to invalid estimates of treatment effect, yet imperfect but plausible choices lead to small biases of the same order of magnitude for all steps. For instance, despite the emphasis often put on choice of confounders, minor deviations from the expert's causal graph did not introduce substantial bias (S6 Appendix), no larger than a too rigid choice of estimator.

To assert the validity of the analysis, we argue to relate as much as possible the average effect to a reference target trial, even when the goal is to capture the heterogeneity of the effect to individualize decisions. EHRs complement RCTs: RCTs cannot address all the subpopulations and local practices [19, 91]. EHRs often cover many individuals, with the diversity needed to model treatment heterogeneity. The corresponding model can then inform better decision-making [17]: a sub-population analysis (as in Figure 4) can distill rules on which groups of patients should receive a treatment. Beyond a sub-group perspective, patient-specific estimates facilitate a personalized approach to clinical decision-making [92].

Even without considering a specific intervention, anchoring machine-learning models on causal mechanisms can make them more robust to distributional shift [93], thus safer and fairer for clinical use [18, 94]. Yet it is important to keep in mind that better prediction is not per se a goal in healthcare. Establishing strong predictors might be less important than identifying moderately strong but modifiable risk factors as established in the Framingham cohort [95], while others such as predictive score are used to optimize population-wide cost-effectiveness, but not individual treatment effect.

No sophisticated data-processing tool can safeguard against invalid study design or a major missing confounder, loopholes that can undermine decision-making systems. Our framework helps the investigator ensure causal validity by outlining the important steps and relating average effects to RCTs. Causal grounding of individual predictions should reduce the social disparities that they reinforce [6, 96, 97], as these are driven by historical decisions and not biological mechanisms. At the population level, it leads to better public health decisions. For instance, going back to cardio-vascular diseases, the stakes are to go beyond risk scores and also account for responder status when prescribing prevention drugs.

Availability of data and materials

The datasets are available on PhysioNet (https://doi.org/10.13026/6mm1-ek67). We used MIMIC-IV.v2.2 The code for data preprocessing and analyses are available on github https://github.com/soda-inria/causal_ehr_mimic/. The 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.

Authors contributions

MD and TS designed the study, MD performed the analysis and wrote the manuscript. TS, JA, CM, LAC, GV reviewed and edited the manuscript.

Acknowledgments

We thank all the PhysioNet team for their encouragements and support. In particular: Fredrik Willumsen Haug, João Matos, Luis Nakayama, Sicheng Hao, Alistair Johnson.

References

1. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ digital medicine. 2018;1(1):18.

2. Liu X, Faes L, Kale AU, Wagner SK, Fu DJ, Bruynseels A, et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. The lancet digital health. 2019;1(6):e271–e297.

3. Li Y, Rao S, Solares JRA, Hassaine A, Ramakrishnan R, Canoy D, et al. BEHRT: transformer for electronic health records. Scientific reports. 2020;10(1):1–12.

4. Beaulieu-Jones BK, Yuan W, Brat GA, Beam AL, Weber G, Ruffin M, et al. Machine learning for patient risk stratification: standing on, or looking over, the shoulders of clinicians? NPJ digital medicine. 2021;4(1):62.

5. Aggarwal R, Sounderajah V, Martin G, Ting DS, Karthikesalingam A, King D, et al. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ digital medicine. 2021;4(1):65.

6. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Annals of internal medicine. 2018;169(12):866–872.

7. Singh H, Mhasawade V, Chunara R. Generalizability challenges of mortality risk prediction models: A retrospective analysis on a multi-center database. PLOS Digital Health. 2022;1(4):e0000023.

8. Gichoya JW, Banerjee I, Bhimireddy AR, Burns JL, Celi LA, Chen LC, et al. AI recognition of patient race in medical imaging: a modelling study. The Lancet Digital Health. 2022;4(6):e406–e414.

9. Seyyed-Kalantari L, Zhang H, McDermott MB, Chen IY, Ghassemi M. Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations. Nature medicine. 2021;27(12):2176–2182.

10. Geirhos R, Jacobsen JH, Michaelis C, Zemel R, Brendel W, Bethge M, et al. Shortcut learning in deep neural networks. Nature Machine Intelligence. 2020;2(11):665–673.

11. Winkler JK, Fink C, Toberer F, Enk A, Deinlein T, Hofmann-Wellenhof R, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA dermatology. 2019;155(10):1135–1141.

12. DeGrave AJ, Janizek JD, Lee SI. AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence. 2021;3(7):610–619.

13. Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. NPJ digital medicine. 2019;2(1):31.

14. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science. 2019;366(6464):447–453.

15. Yuan W, Beaulieu-Jones BK, Yu KH, Lipnick SL, Palmer N, Loscalzo J, et al. Temporal bias in case-control design: preventing reliable predictions of the future. Nature communications. 2021;12(1):1107.

16. Wong A, Otles E, Donnelly JP, Krumm A, McCullough J, DeTroyer-Cooley O, et al. External validation of a widely implemented proprietary sepsis prediction model in hospitalized patients. JAMA Internal Medicine. 2021;181(8):1065–1070.

17. Prosperi M, Guo Y, Sperrin M, Koopman JS, Min JS, He X, et al. Causal inference and counterfactual prediction in machine learning for actionable healthcare. Nature Machine Intelligence. 2020;2(7):369–375.

18. Plecko D, Bareinboim E. Causal fairness analysis. arXiv preprint arXiv:220711385. 2022;.

19. Travers J, Marsh S, Williams M, Weatherall M, Caldwell B, Shirtcliffe P, et al. External validity of randomised controlled trials in asthma: to whom do the results of the trials apply? Thorax. 2007;62(3):219–223.

20. Averitt AJ, Weng C, Ryan P, Perotte A. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. NPJ digital medicine. 2020;3(1):67.

21. Desai RJ, Matheny ME, Johnson K, Marsolo K, Curtis LH, Nelson JC, et al. Broadening the reach of the FDA Sentinel System: a roadmap for integrating electronic health record data in a causal analysis framework. NPJ digital medicine. 2021;4(1):170.

22. Rekkas A, van Klaveren D, Ryan PB, Steyerberg EW, Kent DM, Rijnbeek PR. A standardized framework for risk-based assessment of treatment effect heterogeneity in observational healthcare databases. npj Digital Medicine. 2023;6(1):58.

23. Hernan MA, Sauer BC, Hernandez-Diaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. Journal of clinical epidemiology. 2016;79:70–75.

24. Eriksen CU, Rotar O, Toft U, Jørgensen T. What is the effectiveness of systematic population-level screening programmes for reducing the burden of cardiovascular diseases? World Health Organization. Regional Office for Europe; 2021.

25. Krogsbøll LT, Jørgensen KJ, Larsen CG, Gøtzsche PC. General health checks in adults for reducing morbidity and mortality from disease: Cochrane systematic review and meta-analysis. Bmj. 2012;345.

26. van Staa TP, Smeeth L, Ng ES, Goldacre B, Gulliford M. The efficiency of cardiovascular risk assessment: do the right patients get statin treatment? Heart. 2013;99(21):1597–1602.

27. Krska J, du Plessis R, Chellaswamy H. Implementation of NHS Health Checks in general practice: variation in delivery between practices and practitioners. Primary health care research & development. 2016;17(4):385–392.

28. Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. The Lancet. 2007;370(9596):1453–1457.

29. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS medicine. 2015;12(10):e1001885.

30. Hernàn MA, Robins JM. Causal inference: What If.; 2020.

31. Schneeweiss S, Patorno E. Conducting real-world evidence studies on the clinical outcomes of diabetes treatments. Endocrine Reviews. 2021;42(5):658–690.

32. Zeng J, Gensheimer MF, Rubin DL, Athey S, Shachter RD. Uncovering interpretable potential confounders in electronic medical records. Nature Communications. 2022;13(1):1014.

33. Suissa S. Immortal time bias in pharmacoepidemiology. American journal of epidemiology. 2008;167(4):492–499.

34. Oke J, Fanshawe T, Nunan D. Lead time bias, Catalogue of Bias Collaboration.; 2021. Available from: https://catalogofbias.org/biases/lead-time-bias/.

35. Fu EL, Evans M, Carrero JJ, Putter H, Clase CM, Caskey FJ, et al. Timing of dialysis initiation to reduce mortality and cardiovascular events in advanced chronic kidney disease: nationwide cohort study. bmj. 2021;375.

36. Wang SV, Sreedhara SK, Bessette LG, Schneeweiss S. Understanding variation in the results of real-world evidence studies that seem to address the same question. Journal of Clinical Epidemiology. 2022;151:161–170.

37. Bankhead C ND Aronson JK. Attrition bias, Catalogue of Bias Collaboration.; 2017. Available from: https://catalogofbias.org/biases/attrition-bias/.

38. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999; p. 37–48.

39. VanderWeele TJ. Principles of confounder selection. European journal of epidemiology. 2019;34:211–219.

40. Loh WW, Vansteelandt S. Confounder selection strategies targeting stable treatment effect estimators. Statistics in Medicine. 2021;40(3):607–630.

41. Wang SV, Schneeweiss S, Franklin JM, Desai RJ, Feldman W, Garry EM, et al. Emulation of randomized clinical trials with nonrandomized database analyses: results of 32 clinical trials. JAMA. 2023;329(16):1376–1385.

42. Belloni A, Chernozhukov V, Hansen C. High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives. 2014;28(2):29–50.

43. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, et al.. Double/debiased machine learning for treatment and structural parameters; 2018.

44. Shalit U, Sontag D. Causal Inference for Observational studies: Tutorial; 2016. Available from: https://docplayer.net/64797211-Causal-inference-for-observational-studies.html.

45. Sharma A. Tutorial on causal inference and counterfactual reasoning; 2018. Available from: https://causalinference.gitlab.io/kdd-tutorial/.

46. Moraffah R, Sheth P, Karami M, Bhattacharya A, Wang Q, Tahir A, et al. Causal inference for time series analysis: Problems, methods and evaluation. Knowledge and Information Systems. 2021;63:3041–3085.

47. Stuart EA. Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics. 2010;25(1):1.

48. Austin PC, Stuart EA. Moving towards best practice when using inverse probability of treatment weighting (IPTW) using the propensity score to estimate causal treatment effects in observational studies. Statistics in medicine. 2015;34(28):3661–3679.

49. Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. American Journal of Epidemiology. 1986;123(3):392–402.

50. Johansson FD, Shalit U, Kallus N, Sontag D. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. The Journal of Machine Learning Research. 2022;23(1):7489–7538.

51. Schuler MS, Rose S. Targeted maximum likelihood estimation for causal inference in observational studies. American journal of epidemiology. 2017;185(1):65–73.

52. Dorie V, Hill J, Shalit U, Scott M, Cervone D. Automated versus Do-It-Yourself Methods for Causal Inference. Statistical Science. 2019;34(1):43–68.

53. Alaa A, Van Der Schaar M. Validating causal inference models via influence functions. In: International Conference on Machine Learning. PMLR; 2019. p. 191–201.

54. Curth A, Svensson D, Weatherall J, van der Schaar M. Really doing great at estimating CATE? a critical look at ML benchmarking practices in treatment effect estimation. In: Thirty-fifth conference on neural information processing systems datasets and benchmarks track (round 2); 2021.

55. Johnson A, Bulgarelli L, Pollard T, Horng S, Celi LA, Mark R. Mimic-iv. PhysioNet Available online at: https://physionet org/content/mimiciv/10/(accessed August 23, 2021). 2020;.

56. Hernan MA. Methods of public health research–strengthening causal inference from observational data. New England Journal of Medicine. 2021;385(15):1345–1348.

57. Richardson WS, Wilson MC, Nishikawa J, Hayward RS, et al. The well-built clinical question: a key to evidence-based decisions. Acp j club. 1995;123(3):A12–A13.

58. Riva JJ, Malik KM, Burnie SJ, Endicott AR, Busse JW. What is your research question? An introduction to the PICOT format for clinicians. The Journal of the Canadian Chiropractic Association. 2012;56(3):167.

59. Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. American Journal of Epidemiology. 2016;183(8).

60. Catalogue of Bias Collaboration; 2023. Available from: https://catalogofbias.org/biases/.

61. Weiskopf NG, Dorr DA, Jackson C, Lehmann HP, Thompson CA. Healthcare utilization is a collider: an introduction to collider bias in EHR data reuse. Journal of the American Medical Informatics Association. 2023; p. ocad013.

62. OHDSI. The Book of OHDSI: Observational Health Data Sciences and Informatics. OHDSI; 2021. Available from: https://ohdsi.github.io/TheBookOfOhdsi/.

63. Pearl J, Mackenzie D. The book of why: the new science of cause and effect. Basic books; 2018.

64. Rubin DB. Causal inference using potential outcomes: Design, modeling, decisions. Journal of the American Statistical Association. 2005;100(469):322–331.

65. Attia J, Holliday E, Oldmeadow C. A proposal for capturing interaction and effect modification using DAGs; 2022.

66. Textor J, Hardt J, Knüppel S. DAGitty: a graphical tool for analyzing causal diagrams. Epidemiology. 2011;22(5):745.

67. Cinelli C, Hazlett C. Making sense of sensitivity: Extending omitted variable bias. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2020;82(1):39–67.

68. Tchetgen Tchetgen EJ, Ying A, Cui Y, Shi X, Miao W. An introduction to proximal causal inference. Statistical Science. 2024;39(3):375–390.

69. Wager S. Stats 361: Causal inference; 2020.

70. Imbens GW. Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review. The Review of Economics and Statistics. 2004;86(1):4–29.

71. Colnet B, Josse J, Varoquaux G, Scornet E. Risk ratio, odds ratio, risk difference... Which causal measure is easier to generalize? arXiv preprint arXiv:230316008. 2023;.

72. D'Amour A, Ding P, Feller A, Lei L, Sekhon J. Overlap in observational studies with high-dimensional covariates. Journal of Econometrics. 2021;221(2):644–654.

73. Horng S, Sontag DA, Halpern Y, Jernite Y, Shapiro NI, Nathanson LA. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PloS one. 2017;12(4):e0174708.

74. Jiang LY, Liu XC, Nejatian NP, Nasir-Moin M, Wang D, Abidin A, et al. Health system-scale language models are all-purpose prediction engines. Nature. 2023; p. 1–6.

75. Wendling T, Jung K, Callahan A, Schuler A, Shah NH, Gallego B. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. Statistics in medicine. 2018;37(23):3309–3324.

76. Van der Laan MJ, Polley EC, Hubbard AE. Super learner. Statistical applications in genetics and molecular biology. 2007;6(1).

77. Doutreligne M, Varoquaux G. How to select predictive models for causal inference? arXiv preprint arXiv:230200370. 2023;.

78. Schneeweiss S. Sensitivity analysis and external adjustment for unmeasured confounders in epidemiologic database studies of therapeutics. Pharmacoepidemiology and drug safety. 2006;15(5):291–303.

79. Thabane L, Mbuagbaw L, Zhang S, Samaan Z, Marcucci M, Ye C, et al. A tutorial on sensitivity analyses in clinical trials: the what, why, when and how. BMC medical research methodology. 2013;13(1):1–12.

80. FDA. Statistical Principles for Clinical Trials: Addendum: Estimands and Sensitivity Analysis in Clinical Trials. FDA; 2021.

81. Patel CJ, Burford B, Ioannidis JP. Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. Journal of clinical epidemiology. 2015;68(9):1046–1058.

82. Robertson SE, Leith A, Schmid CH, Dahabreh IJ. Assessing heterogeneity of treatment effects in observational studies. American Journal of Epidemiology. 2021;190(6):1088–1100.

83. Annane D, Siami S, Jaber S, Martin C, Elatrous S, Declere AD, et al. Effects of fluid resuscitation with colloids vs crystalloids on mortality in critically ill patients presenting with hypovolemic shock: the CRISTAL randomized trial. Jama. 2013;310(17):1809–1817.

84. Xu JY, Chen QH, Xie JF, Pan C, Liu SQ, Huang LW, et al. Comparison of the effects of albumin and crystalloid on mortality in adult patients with severe sepsis and septic shock: a meta-analysis of randomized clinical trials. Critical Care. 2014;18(6):1–8.

85. Li B, Zhao H, Zhang J, Yan Q, Li T, Liu L. Resuscitation fluids in septic shock: a network meta-analysis of randomized controlled trials. Shock. 2020;53(6):679–685.

86. Caironi P, Tognoni G, Masson S, Fumagalli R, Pesenti A, Romero M, et al. Albumin replacement in patients with severe sepsis or septic shock. New England Journal of Medicine. 2014;370(15):1412–1421.

87. Zhou S, Zeng Z, Wei H, Sha T, An S. Early combination of albumin with crystalloids administration might be beneficial for the survival of septic patients: a retrospective analysis from MIMIC-IV database. Annals of intensive care. 2021;11:1–10.

88. Battocchi K, Dillon E, Hei M, Lewis G, Oka P, Oprescu M, et al.. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation; 2019. Available from: https://github.com/py-why/EconML.

89. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine learning in Python. the Journal of machine Learning research. 2011;12:2825–2830.

90. Naimi AI, Mishler AE, Kennedy EH. Challenges in obtaining valid causal effect estimates with machine learning algorithms. American Journal of Epidemiology. 2023;192(9):1536–1544.

91. Kennedy-Martin T, Curtis S, Faries D, Robinson S, Johnston J. A literature review on the representativeness of randomized controlled trial samples and implications for the external validity of trial results. Trials. 2015;16:1–14.

92. Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. Bmj. 2018;363.

93. Schölkopf B, Locatello F, Bauer S, Ke NR, Kalchbrenner N, Goyal A, et al. Toward causal representation learning. Proceedings of the IEEE. 2021;109(5):612–634.

94. Richens JG, Lee CM, Johri S. Improving the accuracy of medical diagnosis with causal machine learning. Nature communications. 2020;11(1):3923.

95. Brand RJ, Rosenman RH, Sholtz RI, Friedman M. Multivariate prediction of coronary heart disease in the Western Collaborative Group Study compared to the findings of the Framingham study. Circulation. 1976;53(2):348–355.

96. Mitra N, Roy J, Small D. The Future of Causal Inference. American Journal of Epidemiology. 2022;191(10):1671–1676.

97. Ehrmann DE, Joshi S, Goodfellow SD, Mazwi ML, Eytan D. Making machine learning matter to clinicians: model actionability in medical decision-making. NPJ Digital Medicine. 2023;6(1):7.

98. working group S. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. European heart journal. 2021;42(25):2439–2454.

99. Bretthauer M, Kalager M. Principles, effectiveness and caveats in screening for cancer. Journal of British Surgery. 2013;100(1):55–65.

100. Lee H, Nunan D. Immortal time bias, Catalogue of Bias Collaboration.; 2020. Available from: https://catalogofbias.org/biases/immortaltimebias/.

101. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika;70:41–55.

102. Chernozhukov V, Cinelli C, Newey W, Sharma A, Syrgkanis V. Long story short: Omitted variable bias in causal machine learning. National Bureau of Economic Research; 2022.

103. Jesson A, Mindermann S, Shalit U, Gal Y. Identifying causal-effect inference failure with uncertainty-aware models. Advances in Neural Information Processing Systems. 2020;33:11637–11649.

104. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. American journal of epidemiology. 2011;173(7):731–738.

105. Abadie A, Imbens GW. On the failure of the bootstrap for matching estimators. Econometrica. 2008;76(6):1537–1557.

106. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American statistical Association. 1994;89(427):846–866.

107. Foster DJ, Syrgkanis V. Orthogonal statistical learning. arXiv preprint arXiv:190109036. 2019;.

108. Nie X, Wager S. Quasi-oracle estimation of heterogeneous treatment effects. Biometrika. 2021;108(2):299–319.

109. Robinson PM. Root-N-consistent semiparametric regression. Econometrica: Journal of the Econometric Society. 1988; p. 931–954.

110. Sharma A, Kiciman E. DoWhy: An end-to-end library for causal inference. arXiv preprint arXiv:201104216. 2020;.

111. Bouthillier X, Delaunay P, Bronzi M, Trofimov A, Nichyporuk B, Szeto J, et al. Accounting for variance in machine learning benchmarks. Proceedings of Machine Learning and Systems. 2021;3:747–769.

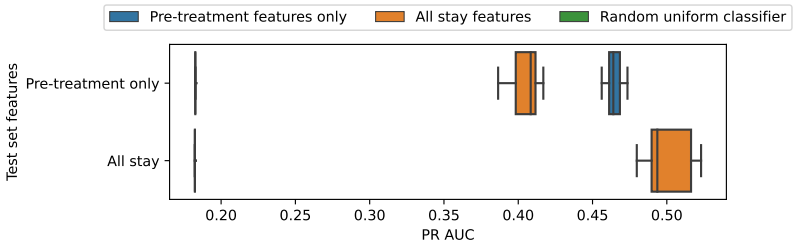
112. Investigators SS. Saline or albumin for fluid resuscitation in patients with traumatic brain injury. New England Journal of Medicine. 2007;357(9):874–884.

Supporting information

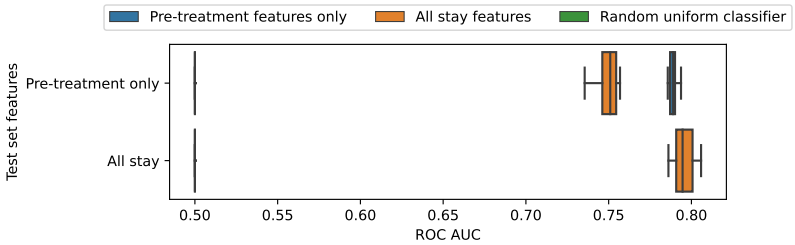
S1 Fig. Motivating example: Failure of predictive models to predict mortality from pretreatment variables. To illustrate how machine learning frameworks can fail to inform decision making, we present a motivating example from MIMIC-IV. Using the same population and covariates as in the main analysis (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 [98]: *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 [99].

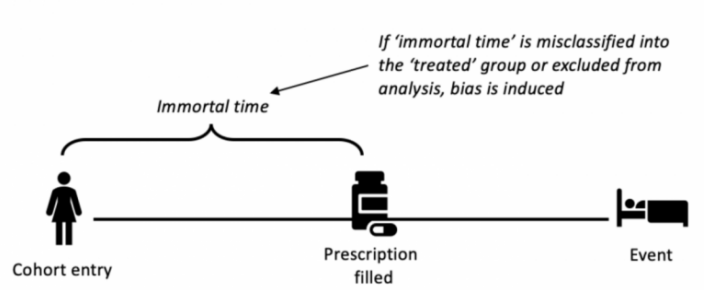


Fig 6. *Poor experimental design can introduce Immortal time bias, which leads to a treated group with falsely longer longevity [100].*

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.

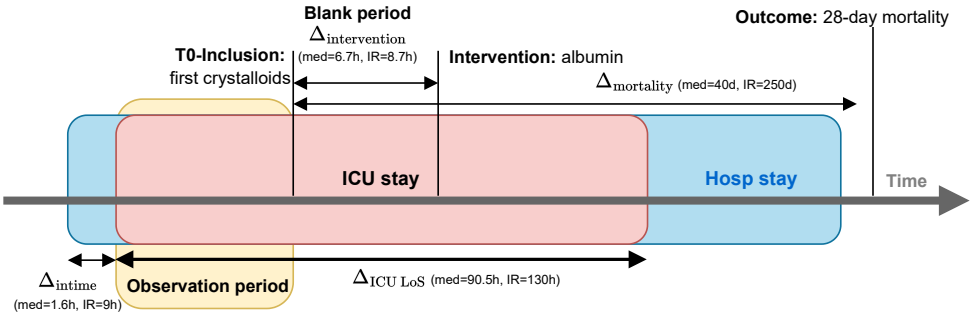


Fig 7. *Defining the inclusion event, the starting time T0 for follow-up, the intervention's assignment time and the observation window for confounders is crucial to avoid time and selection biases. In our study, the gap between the intervention and the inclusion is small compared to the occurrence of 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.

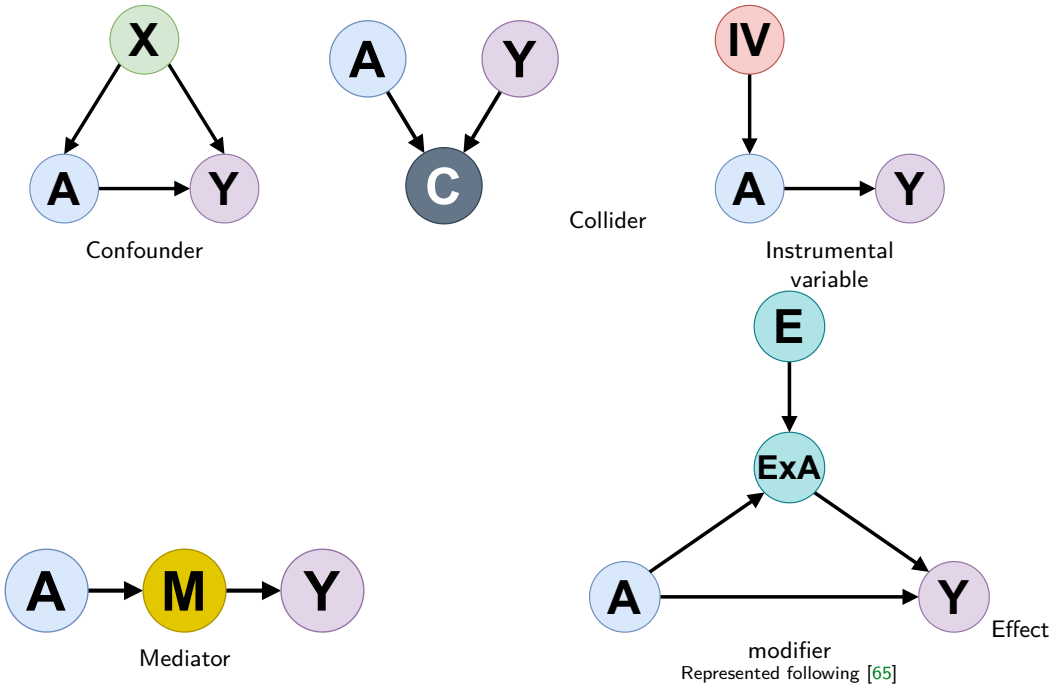


Fig 8. *The five categories of causal variables needed for our framework: A: Treatment, X: Confounder, IV: Instrumental variable, M: mediator, Y: Outcome, C: Collider, E: Effect modifier.*

S1 Appendix. 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 [64], 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 $e(X) = \mathbb{P}(A = 1|X)$ *[101]:* $\{Y(0), Y(1)\} \perp\!\!\!\perp A|e(X)$.

Unconfoundedness is a strong assumption that might be violated in practice. The existance of residual bias trough unobserved confounders can be mitigated with different strategies. The *omitted variable bias* framework encourages sensitivity analyses allowing to derive bounds on the causal estimate by making assumptions on the strengh of association of the omitted variable with both the treatment and the outcome. We refer to [67] for a clear introduction under linear assumption and to [102] for an extension to general non-linear settings. In case of strong unobserved confounders for which proxy variables can be measured, *proximal inference* can be used to obtain identifiability [68]. These methods require expert knowledge to classify the proxy between treatment and outcome proxy, then run two-stage regression to recover the causal effect. Lastly, natural exeperiments, when available, should be exploited to estimate causal effects without the need of unconfoundedness. Instrumental variable methods exploit randomness influancing the treatment but unrelated to the outcome to simulate a randomized experiment [69, chapter 9]. Regression discontinuity designs leverage discontinuous treatment assignment mechanisms with the assumption of a continuous outcome [69, chapter 5].

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 [72], 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 [30].

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 [103].*

S2 Appendix. Major causal-inference methods: When to use which estimator?

G-formula also called conditional mean regression [75], g-computation [49], or Q-model [104]. This approach is directly modeling the outcome, also referred to as the response surface: $\mu_{(a)}(x) = \mathbb{E}\left(Y \mid A = a, \mathbf{X} = x\right)$

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 [47]. However, the number of confounder bins grows exponentially with the number of variables. [101] 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 [105] 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 [48]. 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) [106].

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:

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

Moreover, despite the need to estimate two models, this estimator is more efficient in the sense that it converges quicker than single model estimators [69]. 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, [107] 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 [43] also known as the R-learner [108]. It is based on the R-decomposition, [109], 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:

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

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.*

S3 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 [89] 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 [69] 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

S4 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 [88] 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 [110] 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*

S5 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 [111, 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

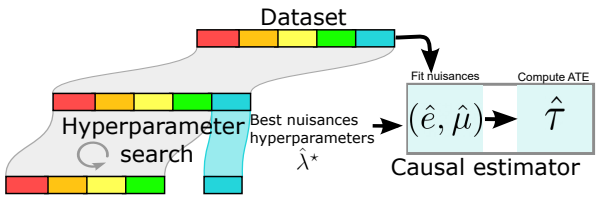


Fig 9. *Hyper-parameter search procedure.*

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.

S5 Fig. Selection flowchart.

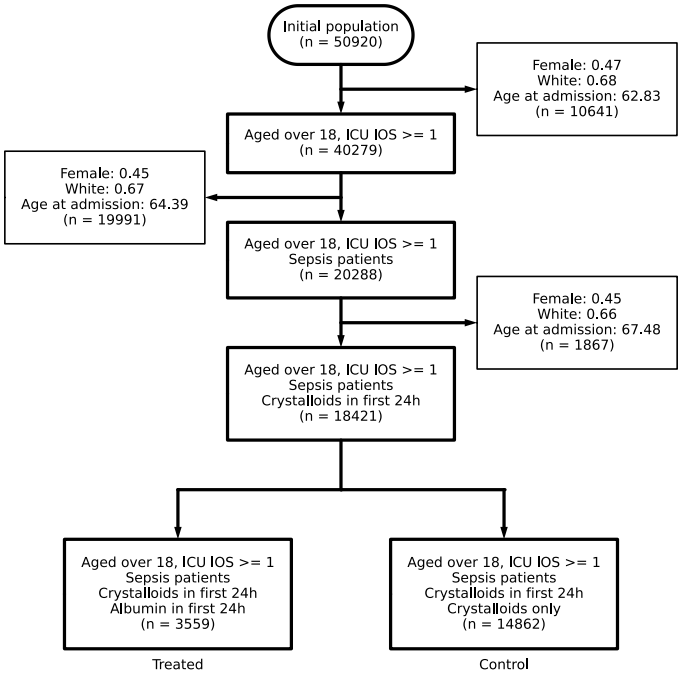


Fig 10. *Selection flowchart on MIMIC-IV for the emulated trial.*

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.

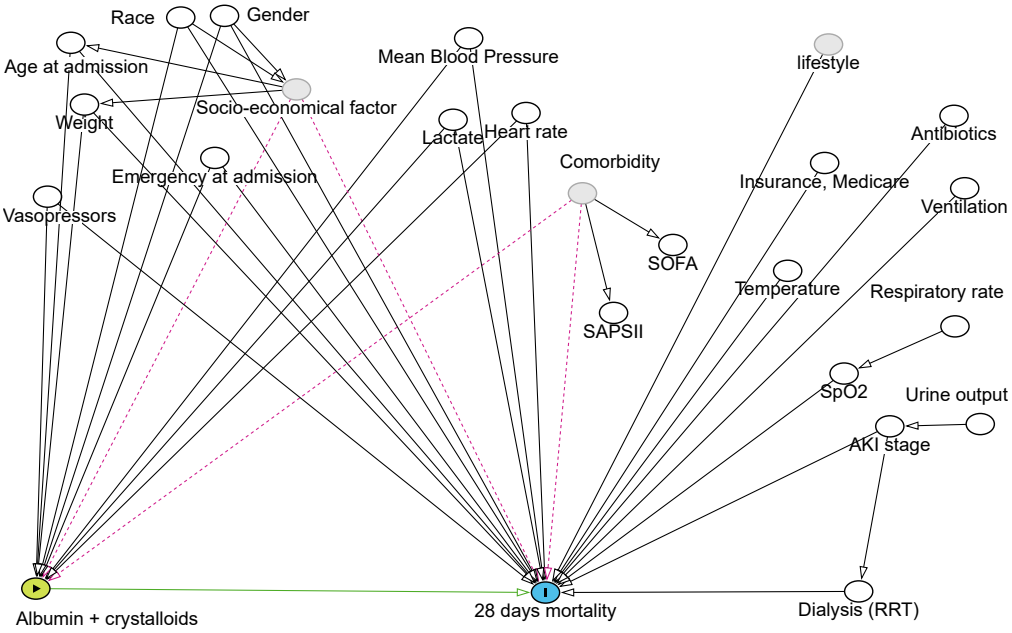


Fig 11. ***Causal graph for the Albumin vs crystalloids emulated trial*** – *The green arrow indicates the effect studied. Black arrows show causal links known to medical expertise. Dotted red arrows highlight confounders not directly observed. For readability, we draw only the most important edges from 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).

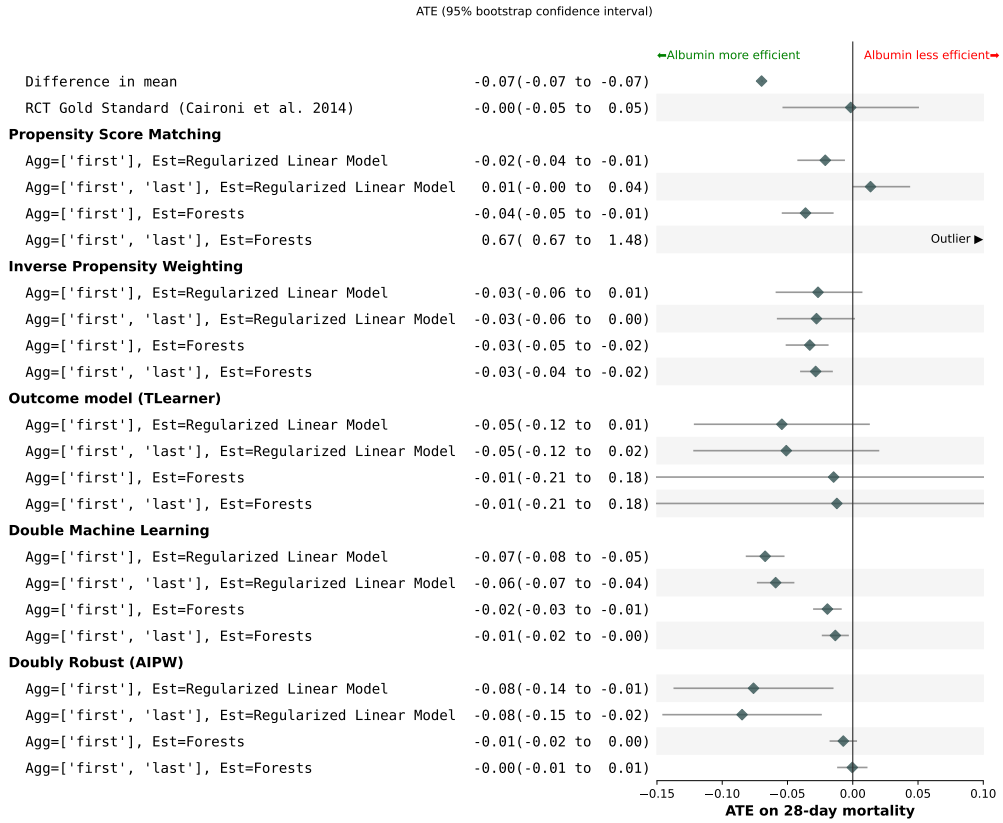


Fig 12. *Full sensitivity analysis: The estimators with forest nuisances point to no effect for almost every causal estimator consistently with the RCT gold standard. Only matching with forest yields an unconvincingly high estimate. Linear nuisance used with doubly robust methods suggest a reduced mortality 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.

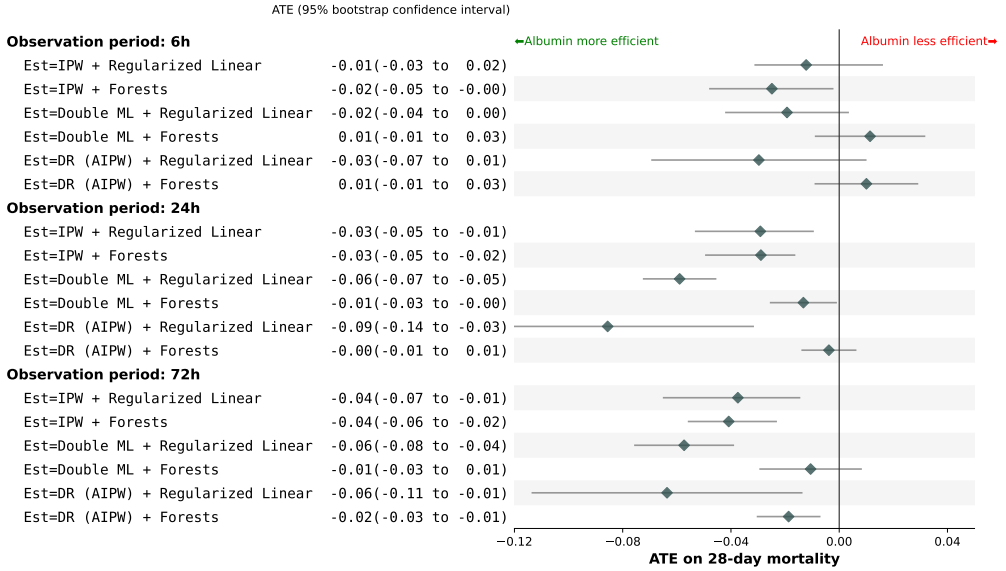


Fig 13. *Sensitivity analysis for immortal time bias: Every choice of estimates show an improvement of the albumin treatment when increasing the observation period, thus increasing the blank period between inclusion and administration of albumin. Aggregation was concatenation of first and last features.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

S6 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).

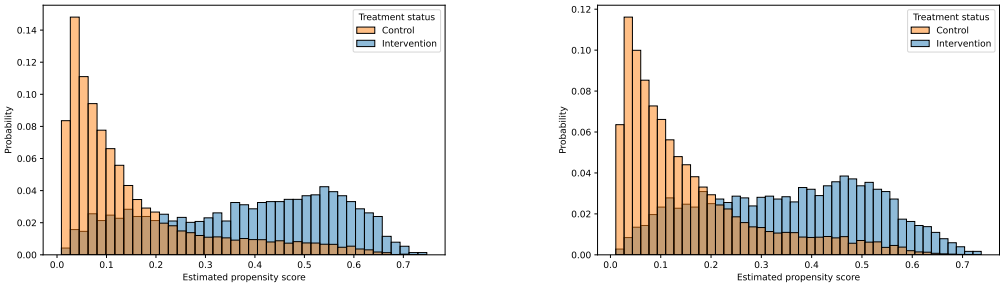
Figure 3b shows that small deviation from the ignorability assumptions is tolerable: for example, removing antibiotics does not impact the estimate. However, the larger the deviation from graph S6 Fig, the larger the bias compared to the gold-standard. Adjusting only for socio-demographics features is 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 [40].

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 [48], 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 [77]. 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.

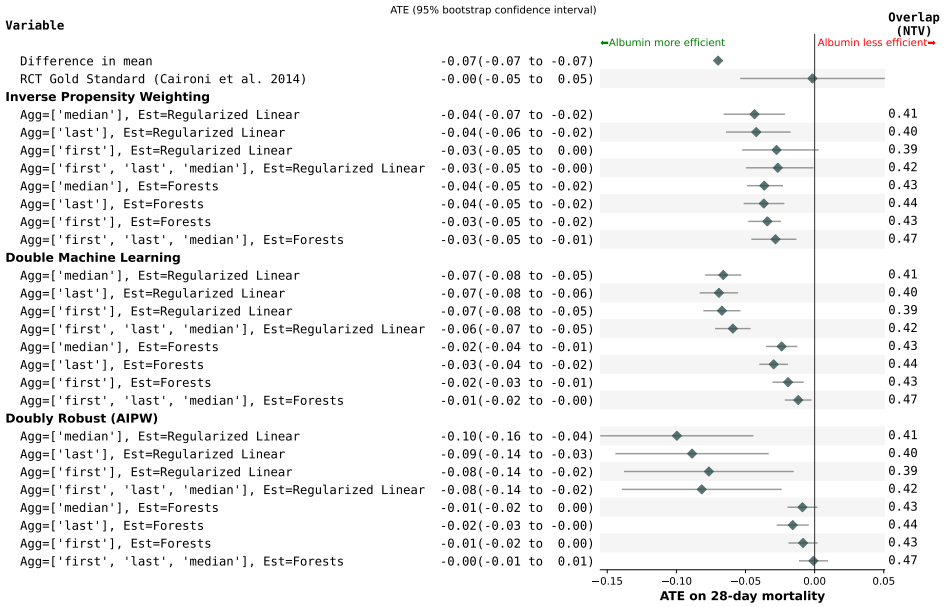


Fig 15. *Vibration analysis dedicated to the aggregation choices. The choices of aggregation only marginally modify the results. When assessed with Normalized Total Variation, the overlap assumption is respected for all our choices of aggregation. The green diamonds depict the mean effect and the bar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

S7 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

[86] 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).

[112] 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.

[87] 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.

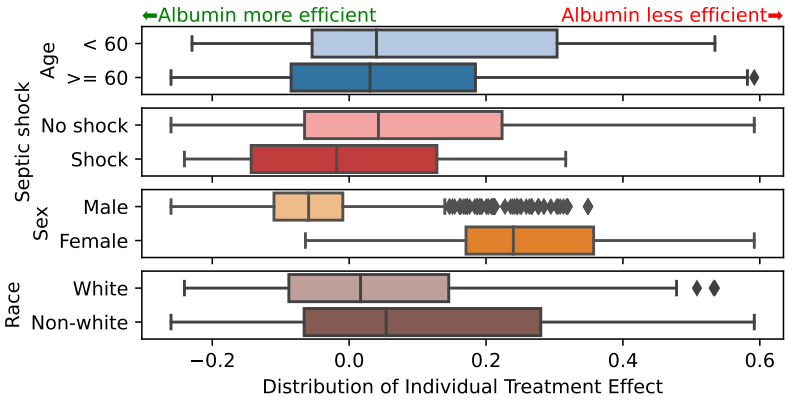


Fig 16. *Distribution of Conditional Average Treatment effects on sex, age, race and pre-treatment septic shock estimated with a final forest estimator. The CATE are positive for each subgroups, which is not consistent with the null treatment effect obtained in the main analysis. The boxes contain between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

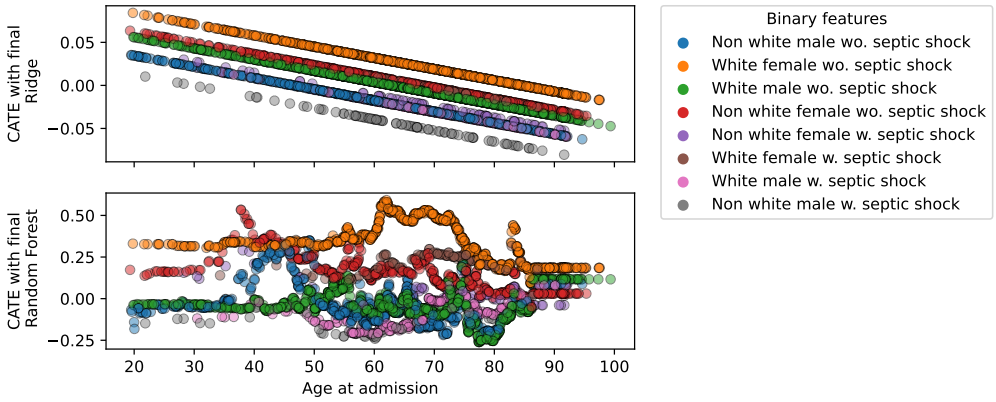


Fig 17. *Distribution of Conditional Average Treatment effects on sex, age, race and pre-treatment septic shock plotted for different ages. On the top the final estimator is a linear model; on the bottom, it is a random forest. The forest-based CATE displays more noisy trends than the linear-based CATE.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

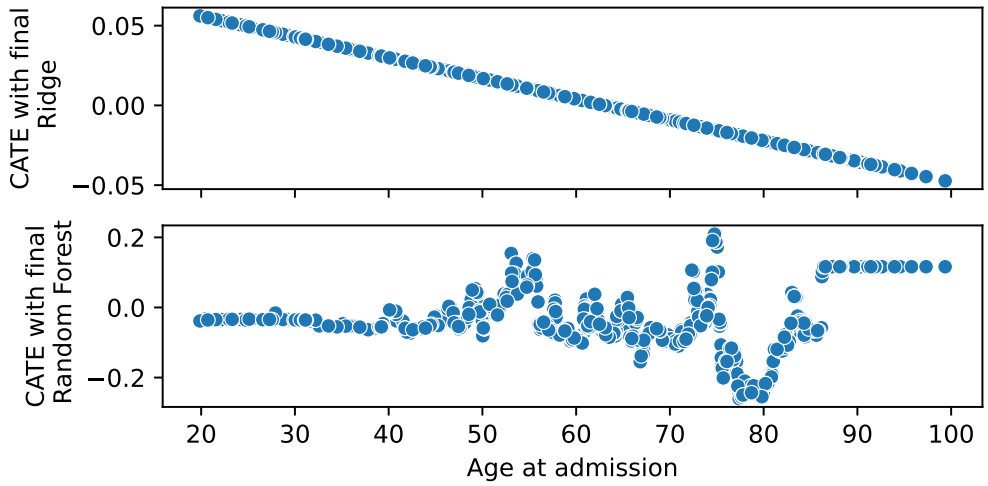


Fig 18. *Figure 17 on the subpopulation of white male patients without septic shock. Contrary to the ridge regression (on top) inducing a nicely interpretable trend, using random forests as the final estimator failed to recover CATE on ages: the predicted estimates do not exhibit any trend and display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