Placeholder for OUP logo oup.pdf

GigaScience, 2024, 1–28

doi: xx.xxxx/xxxx
Manuscript in Preparation
Paper

PAPER

How to select predictive models for decision-making or causal inference?

Matthieu Doutreligne[1,*] and Gaël Varoquaux[2]

[1]Soda, Inria Saclay, France and [2]Mission Data, Haute Autorité de Santé, France

[*]matt.dout@gmail.com; matthieu.doutreligne@inria.fr

Abstract

Background: We investigate which procedure selects the predictive model most trustworthy to reason on the effect of an intervention and support decision making.
Methods: We study ~~such a~~ a large variety of model selection procedures in practical settings: finite samples settings and without theoretical assumption of *well-specified* models. Beyond standard cross-validation or internal validation procedures, we also study elaborate causal risks. These build proxies of the causal error using "nuisance" re-weighting to compute it on the observed data. We evaluate whether empirically estimated nuisances, which are necessarily noisy, add noise to model selection. We compare different metrics for causal model selection in an extensive empirical study based on a simulation and three healthcare datasets based on real covariates.
Results: Among all metrics, the mean squared error, classically used to evaluate predictive modes, is worse. Re-weighting it with propensity score does not bring much improvements ~~. The~~ in most cases. On average, the R-risk, which uses as nuisances a model of mean outcome and propensity scores, leads to the best performances. Nuisance corrections are best estimated with flexible estimators such as a super learner.
Conclusions: When predictive models are used to reason on the effect of an intervention, they must be evaluated with different procedures than standard predictive settings; using the R-risk from causal inference.

Key words: Model Selection, Predictive model, Treatment Effect, ~~G-formula~~G-computation, Machine Learning

Introduction

Extending prediction to prescription needs causality

Prediction models have long been used in biomedical settings, as with risk score or prognostic models [1, 2]. While these have historically been simple models on simple data, this is changing with progress in machine learning and richer medical data [3, 4]. Health predictions can now integrate medical images [5, 6, 7, 8, 9], patient records [10, 11, 12] or clinical notes [13, 14, 15]. Complex data is difficult to control and model, but these models are validated by verifying the accuracy of the prediction on left-out data [16, 17, 18]. Crucial to the clinical adoption of a model predicting a health outcome is that it "can support decisions about patient care" [19]. Precision medicine is about guiding decisions: *eg* will an individual benefit from an intervention such as surgery [20]? An estimate of the effect of the treatment can be obtained by contrasting model predictions with and without the treatment, but statistical validity requires causal inference [21, 22, 23].

Indeed, concluding on the effect of a treatment is a difficult causal-inference task, as it can be easily compromised by confounding: spurious associations between treatment allocation and baseline health, *e.g.* only prescribing a drug to mild cases [24, 25]. Predictive modeling ~~bridges~~ is linked to causal inference theory ~~under the name~~ by the concept of *outcome models* (or ~~G-computation, G-formula~~ g-computation, g-estimation, g-formula [26], Q-model [21], conditional mean regression [27]). Medical statistics and epidemiology have mostly used other causal-inference methods, modeling treatment assignment with propensity scores [28, 29, 30, 31]. Outcome modeling brings the benefit of going beyond average effects, estimating individualized or conditional average treatment effects (CATE), central to precision

Compiled on: October 20, 2024.
Draft manuscript prepared by the author.

medicine. For this purpose, such methods are also invaluable for randomized trials [32, 33, 34].

Outcome-modeling methods, even when specifically designed for causal inference, are numerous: Bayesian Additive Regression Trees [35], Targeted Maximum Likelihood Estimation [36, 37], causal boosting [38], causal multivariate adaptive regression splines [38], random forests [39, 40], Meta-learners [41], R-learners [42], Doubly robust estimation [43]... The wide variety of methods raises the problem of selecting between different estimators based on the data at hand. Indeed, estimates of treatment effects can vary markedly across different predictive models ~~[44, 45]~~ [44, 45, 46, 47] (illustration in Appendix A.1).

Given complex health data, which predictive model is to be most trusted to yield valid causal estimates needed to motivate individual treatment decisions? As no single machine-learning method performs best across all ~~data sets~~datasets, there is a pressing need for clear guidelines to select outcome models for causal inference.

Objectives and structure of the paper. ~~We study~~ The intersection between machine learning and causal inference is growing rapidly [48, 49]. We focus on *model selection procedures* in practical settings, without theoretical assumptions often made in statistical literature such as *infinite* data or *well-specified* models (Appendix A.2). Asymptotic causal-inference theory recommends complex risks, but a practical question is whether model-selection procedures, that rely on data split, can estimate these risks reliably enough. Indeed, these risks come with more quantities to estimate, which may bring additional variance, leading to worse model selection.

We first illustrate the problem of causal model selection. Then we anchor causal model selection in the *potential outcome* framework and ~~details~~ detail the causal risks and model-selection procedure. We then rewrite the so-called R-risk as a reweighted version of mean squared difference between the true and estimated individualized treatment effect. Finally, we conduct a thorough empirical study comparing the different metrics on diverse datasets, using a family of simulations and real health data, going beyond prior work limited to specific simulation settings [50, 51] (Appendix A.2).

Illustration: the best predictor may not estimate best causal effects

Using a predictor to reason on causal effects relies on contrasting the prediction of the outcome for a given individual with and without the treatment. Given various predictors of the outcome, which one should we use? Standard predictive modeling or machine-learning practice selects the predictor that minimizes the expected error on the outcome [17, 18]. However, this predictor may not be the best model to reason about causal effects of an intervention as Figure 1 illustrates. Consider the probability Y of an undesirable outcome (*e.g.*death), a binary treatment $A \in \{0, 1\}$, and a covariate $X \in \mathbb{R}$ summarizing the patient health status (*e.g.*the Charlson index [52]). We simulate a treatment beneficial (decreases mortality) for patients with high Charlson scores (bad health status) but with little effect for patients in good condition (low Charlson scores).

Figure 1a shows a random forest predictor with a counter-intuitive behavior: it predicts well on average the outcome (as measured by a regression R^2 score) but perform poorly to estimate causal quantities: the average treatment effect τ (as visible via the error $|\tau - \hat{\tau}|$) or the conditional average treatment effect (the error $\mathbb{E}[(\tau(x) - \hat{\tau}(x))^2]$, called CATE). On the contrary, Figure 1b shows a linear model with smaller R^2 score but better causal inference.

The problem is that causal estimation requires controlling an error on both treated and non-treated outcome for the same individual: the observed outcome, and the non-observed *counterfactual* one. The linear model is misspecified –the outcome functions are not linear–, leading to poor R^2; but it interpolates better to regions where there are few untreated individuals –high Charlson score–

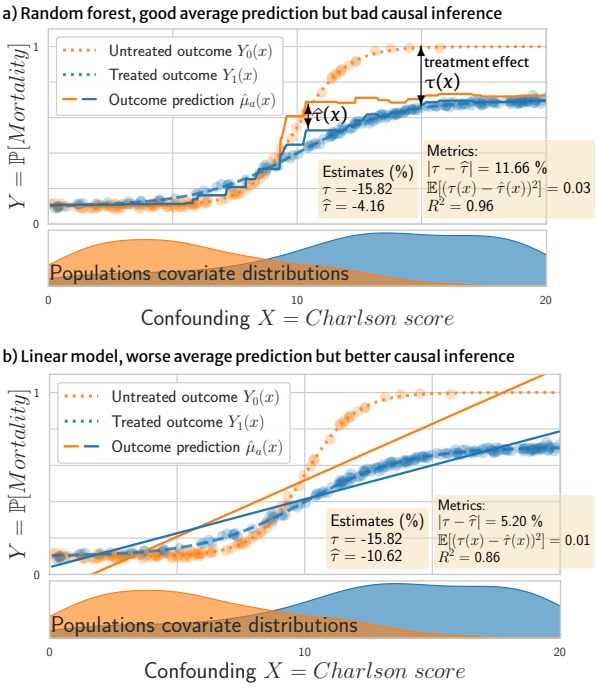


Figure 1. Illustration: a) a random-forest predictor with high performance for standard prediction (high R^2) but that yields poor causal estimates (large error between true effect τ and estimated $\hat{\tau}$), b) a linear predictor with smaller prediction performance leading to better causal estimation.

Selecting the predictor with the smallest error to the individual treatment effect $\mathbb{E}[(\tau(x) - \hat{\tau}(x))^2]$ –the τ-risk, eq. 10 – would lead to the best causal estimates; however computing this error is not feasible: it requires access to unknown quantities: $\tau(x)$.

While the random forest fits the data better than the linear model, it gives worse causal inference because its error is inhomogeneous between treated and untreated. The R^2 score does not capture this inhomogeneity.

and thus gives better causal estimates. Conversely, the random forest puts weaker assumptions on the data, thus has higher R^2 score but is biased by the treated population in the poor-overlap region, leading to bad causal estimates.

This toy example illustrates that the classic minimum ~~Mean Square Error~~ mean squared error (MSE) criterion is not suited to choosing a model among candidate estimators for causal inference.

Methods

Neyman-Rubin Potential Outcomes framework

We first expose the classic construction of the outcome modeling (or G-computation) estimators of causal effect [53, 21, 24].

Settings. The Neyman-Rubin Potential Outcomes framework [54, 55] enables statistical reasoning on causal treatment effects: Given an outcome $Y \in \mathbb{R}$ (*e.g.*mortality risk or hospitalization length), function of a binary treatment $A \in \mathcal{A} = \{0, 1\}$ (*e.g.* a medical procedure), and baseline covariates $X \in \mathcal{X} \subset \mathbb{R}^d$, we observe the factual distribution, $O = (Y(A), X, A) \sim \mathcal{D} = \mathbb{P}(y, x, a)$. However, we want to model the existence of potential observations (unobserved ie. counterfactual) that correspond to a different treatment. Thus we want quantities on the counterfactual distribution $O^* = (Y(1), Y(0), X, A) \sim \mathcal{D}^* = \mathbb{P}(y(1), y(0), x, a)$.

Popular quantities of interest (estimands) are: at the population level, the Average Treatment Effect

$$\text{ATE} \qquad \tau \overset{\text{def}}{=} \mathbb{E}_{Y(1),Y(0)\sim\mathcal{D}^*}[Y(1) - Y(0)];$$

at the individual level, to model heterogeneity, the Conditional Average Treatment Effect

$$\text{CATE} \qquad \tau(x) \overset{\text{def}}{=} \mathbb{E}_{Y(1),Y(0)\sim\mathcal{D}^\star}[Y(1) - Y(0)|X = x].$$

Causal assumptions. A given data needs to meet a few assumptions to enable identifying causal estimands [56]. ~~The usual *strong ignorability* assumptions are (details in A.3): *1)*~~ an individual's outcome Y is solely governed by the corresponding potential outcome:

$$\text{Consistency assumption,} \qquad Y = A\,Y(1) + (1 - A)\,Y(0) \tag{1}$$

2) unconfoundedness $\{Y(0), Y(1)\} \perp\!\!\!\perp A|X$, ~~*3)*~~ *strong overlap* ie. every patient has a strictly positive probability to receive each treatment, ~~*3) consistency,*~~ and *4) generalization* –no covariate shift. These classic assumptions, called *strong ignorability*, are formally detailed in A.3.

~~*Estimating*~~ *Identifying treatment effects with outcome models – g–computation [53].* Should we know the two ~~expected~~ potential outcomes for a given X, we could compute the difference between them, which gives the causal effect of the treatment. These two ~~expected~~ potential outcomes can be estimated from observed data: the consistency 1 and unconfoundedness 2 assumptions imply the following equality, linking the target quantity to the observed data:

$$\mathbb{E}_{Y(a)\sim\mathcal{D}^\star}[Y(a)|X = x] = \mathbb{E}_{Y\sim\mathcal{D}}[Y|X = x, A = a] \tag{2}$$

On the left, the expectation is taken on the counterfactual unobserved distribution. On the right, the expectation is taken on the factual observed distribution conditionally on the treatment. For the rest of the paper, the expectations will always be taken on the factual observed distribution $\mathcal{D}$. This identification leads to outcome based estimators (ie. g-computation estimators [21]):

$$\begin{aligned}\tau = \quad & \mathbb{E}_{Y\sim\mathcal{D}^\star}[Y(1) - Y(0)|X = x]_{Y\sim\mathcal{D}^\star}[Y(1) - Y(0)] \\ = \quad & \mathbb{E}_{Y\sim\mathcal{D}}[Y|A = 1] - \mathbb{E}_{Y\sim\mathcal{D}}[Y|A = 0]\end{aligned} \tag{3}$$

This equation builds on ~~two quantities: the conditional expectancy~~ the conditional expectation of the outcome given the ~~covariates and either treatment or no no treatment~~ treatment $\mathbb{E}_{\sim\mathcal{D}}[Y|A]$. Outcome based methods target this quantity conditionally on the covariates, called *response function*:

$$\text{Response function} \qquad \mu_a(x) \overset{\text{def}}{=} \mathbb{E}_{Y\sim\mathcal{D}}[Y|X = x, A = a]$$

Given a sample of data and the oracle response functions μ_0, μ_1, the finite sum version of Equation 3 leads to an unbiased estimator of the ATE written:

$$\hat{\tau} = \frac{1}{n}\left(\sum_{i=1}^{n} \mu_1(x_i) - \mu_0(x_i)\right) \tag{4}$$

This estimator is an oracle *finite sum* estimator by opposition to the population expression of τ, $\mathbb{E}[\mu_1(x_i) - \mu_0(x_i)]$, which involves an expectation taken on the full distribution $\mathcal{D}$, which is observable but requires infinite data. For each estimator ℓ taking an expectation over $\mathcal{D}$, we use the symbol $\hat{\ell}$ to note its finite sum version. The formulas in Eq. (2–4) are all partly oracle formulas: they rely on conditional expectations, the response functions, but give no specific procedures on how to compute or select them. This last point is the topic of our work, describe in the next section.

Similarly to the ATE, at the individual level ~~, the CATE~~ Eq.2 links the CATE to statistical quantities:

$$\tau(x) = \mu_1(x) - \mu_0(x) \tag{5}$$

Robinson decomposition. ~~The *R-decomposition* of the outcome model[57]plays an important role: introducing two quantities~~G-computation is a choice of decomposition of the CATE estimation. Other choices of decomposition exist, such as the R-decomposition [57]. The latter introduces two new statistical estimates, the conditional mean outcome and the probability ~~to be~~ of being treated (known as propensity score [28]):

$$\text{Conditional mean outcome} \qquad m(x) \overset{\text{def}}{=} \mathbb{E}_{Y\sim\mathcal{D}}[Y|X = x] \tag{6}$$

$$\text{Propensity score} \qquad e(x) \overset{\text{def}}{=} \mathbb{P}[A = 1|X = x] \tag{7}$$

~~the outcome~~ with these, the outcome (Eq. 1) can be written

$$\begin{aligned}\text{R-decomposition} \qquad & y(a) = m(x) + (a - e(x))\tau(x) + \varepsilon(x;a) \\ & \text{with} \quad \mathbb{E}[\varepsilon(X;A)|X,A] = 0\end{aligned} \tag{8}$$

m and e are often called *nuisances* [43]~~; they are unknown~~. They are unknown and must be estimated from the data.

Both the in the ATE and CATE formula Eq. (4, 5), and the Robinson decomposition involve conditional expectations –the response functions $\mu_a(x)$ or the nuisances $m(x)$ and $e(x)$. In practice those are given by statistical models: linear models, random forests, etc [48, 49].

Model-selection~~risks~~, oracle and feasible risks

Causal model selection. We formalize model selection for causal estimation. Thanks to the ~~g–formula~~ outcome model identification (Equation 2), a given ~~outcome~~ model $f : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ –learned from data or built from domain knowledge– induces feasible estimates of the ATE and CATE (eqs 4 and 5), $\hat{\tau}_f$ and $\hat{\tau}_f(x)$. ~~Let~~ However, the g-computation framework presented above is written in terms of "perfect" conditional expectations (oracles), it does not control an error, *eg* on both populations as highlighted in Figure 1. Selection procedures are needed to find the best conditional-expectation models.

A selection procedure combines a risk ℓ, evaluating the quality of a model f with observed data O, and a splitting strategy of the data to estimate different regressions (nuisances) involved in the risk. Formally, let $\mathcal{F} = \{f : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}\}$ be a family of such estimators. Our goal is to select the best candidate in this family for the observed dataset O using a risk ℓ:

$$f^*_\ell = \underset{f\in\mathcal{F}}{\operatorname{argmin}}\, \ell(f, O) \tag{9}$$

We now detail possible risks ℓ, risks useful for causal model selection, and how to compute them.

The τ-risk: an oracle error risk. As we would like to target the CATE, the following evaluation risk is natural (also called PEHE [59, 35]):

$$\tau\text{-risk}(f) \overset{\text{def}}{=} \mathbb{E}_{X\sim p(X)}[(\tau(X) - \hat{\tau}_f(X))^2] \tag{10}$$

Given observed data from $p(X)$, the expectation is computed with a finite sum, as in eq. 4, to give an estimated value $\widehat{\tau\text{-risk}}(f)$. However this risk is not feasible as the oracles $\tau(x)$ are not accessible with the observed data $(Y, X, A) \sim \mathcal{D}$.

Table 1. Review of causal risks — The R-risk* is called τ-risk$_R$ in [50].

Risk	Equation	Reference
τ-risk = MSE$(\tau(X), \tau_f(X))$	$\mathbb{E}_{X\sim p(X)}[(\tau(X) - \hat{\tau}_f(X))^2]$	Eq. 10 [35]
μ-risk = MSE$(Y, f(X))$	$\mathbb{E}_{(Y,X,A)\sim\mathcal{D}}\left[(Y - f(X;A))^2\right]$	Def. 1 [50]
μ-risk$^*_{IPW}$	$\mathbb{E}_{(Y,X,A)\sim\mathcal{D}}\left[\left(\frac{A}{e(X)} + \frac{1-A}{1-e(X)}\right)(Y - f(X;A))^2\right]$	Def. 2 [58]
τ-risk$^*_{IPW}$	$\mathbb{E}_{(Y,X,A)\sim\mathcal{D}}\left[\left(Y\left(\frac{A}{e(X)} - \frac{1-A}{1-e(X)}\right) - \hat{\tau}_f(X)\right)^2\right]$	Def. 3 [39]
U-risk*	$\mathbb{E}_{(Y,X,A)\sim\mathcal{D}}\left[\left(\frac{Y-m(X)}{A-e(X)} - \hat{\tau}_f(X)\right)^2\right]$	Def. 4 [42]
R-risk*	$\mathbb{E}_{(Y,X,A)\sim\mathcal{D}}\left[\left((Y - m(X)) - (A - e(X))\hat{\tau}_f(X)\right)^2\right]$	Def. 5 [42]

Feasible error risks. Table 1 lists *feasible* risks (Detailed in Appendix A.4), based on the prediction error of the outcome model and *observable* quantities. These observable, called nuisances are e – propensity score, eq 7– and m –conditional mean outcome, eq 6. We give the definitions as *semi-oracles*, function of the true unknown nuisances, but later instantiate them with estimated nuisances, noted $(\check{e}, \check{m})$. Semi-oracles risks are superscripted with the * symbol.

~~Estimation and model~~ Model selection procedure

Causal model selection (eq 9) may involve estimating various quantities from the observed data: the outcome model f, its induced risk as introduce in the previous section, and possibly nuisances required by the risk. Given a dataset with N samples, we split out a train and a test sets $(\mathcal{T}, \mathcal{S})$. We fit each candidate estimator $f \in \mathcal{F}$ on $\mathcal{T}$. We also fit the nuisance models $(\check{e}, \check{m})$ on the train set $\mathcal{T}$, setting hyperparameters by a nested cross-validation before fitting the nuisance estimators with these parameters on the full train set. Causal quantities are then computed by applying the fitted candidates estimators $f \in \mathcal{F}$ on the test set $\mathcal{S}$. Finally, we compute the model-selection metrics for each candidate model on the test set. This procedure is described in Algorithm 1 and Figure 2.

Algorithm 1 Model selection procedure

Given train and test sets $(\mathcal{T}, \mathcal{S}) \sim \mathcal{D}$, a candidate ~~estimators~~ estimator f, a causal metrics ℓ:

i. Prefit: Learn estimators for unknown nuisance quantities $(\check{e}, \check{m})$ on the training set $\mathcal{T}$
ii. Fit: learn $\hat{f}(\cdot, a)$ on $\mathcal{T}$
iii. Model selection: $\forall x \in \mathcal{S}$ predict $(\hat{f}(x,1), \hat{f}(x,0))$ and evaluate the estimator storing the metric value: $\ell(f, \mathcal{S})$ – possibly function of $\check{e}$ and $\check{m}$

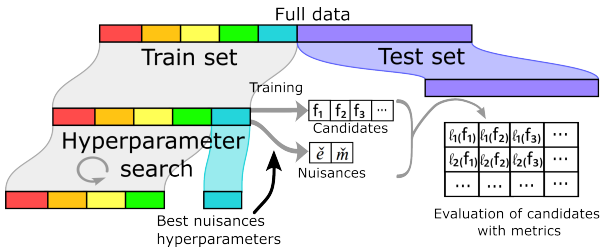


Figure 2. Estimation procedure for causal model selection.

R-risk as reweighted oracle metric

The R-risk can be rewritten as a rebalanced τ-risk.

This rewriting involves reweighted residuals: for each potential outcome, $a \in \{0;1\}$, the variance conditionally on x is [60]:

$$\sigma^2_y(x;a) \stackrel{\text{def}}{=} \int_y (y - \mu_a(x))^2 \, p(y \mid x = x; A = a)\, dy$$

Integrating over the population, we get the Bayes squared error: $\sigma^2_B(a) = \int_{\mathcal{X}} \sigma^2_y(x;a)p(x)dx$ and its propensity weighted version: $\tilde{\sigma}^2_B(a) = \int_{\mathcal{X}} \sigma^2_y(x;a)\, p(x;a)\, dx$. In case of a purely deterministic link between the covariates, the treatment, and the outcome, these residual terms are null.

Proposition 1 (R-risk as reweighted τ-risk) *Given an outcome model f, its R-risk appears as weighted version of its τ-risk (Proof in A.5):*

$$R\text{-risk}^*(f) = \int_x e(x)(1 - e(x))\,(\tau(x) - \tau_f(x))^2 p(x)dx + \tilde{\sigma}^2_B(1) + \tilde{\sigma}^2_B(0) \tag{11}$$

The R-risk targets the oracle at the cost of an overlap reweighting and the addition of the reweighted Bayes residuals, which are independent of f. In good overlap regions the weights $e(x)(1 - e(x))$ are close to $\frac{1}{4}$, hence the R-risk is close to the desired gold-standard τ-risk. For randomized control trials, this weight is constant making the R-risk particularly suited for exploring heterogeneity (Appendix A.5)

Empirical Study

We evaluate the following causal metrics, oracle and feasible versions, presented in Table 1: $\widehat{\mu\text{-risk}^*_{IPW}}$, $\widehat{R\text{-risk}^*}$, $\widehat{U\text{-risk}^*}$, $\widehat{\tau\text{-risk}^*_{IPW}}$, $\widehat{\mu\text{-risk}}$, $\widehat{\mu\text{-risk}_{IPW}}$, $\widehat{R\text{-risk}}$, $\widehat{U\text{-risk}}$, $\widehat{\tau\text{-risk}_{IPW}}$. We benchmark the metrics in a variety of settings: many different simulated data generation processes and three semi-simulated datasets [1].

The simulations have been designed to evaluate the effect of the overlap parameter. The semi-simulated datasets have been included to explore more diverse and noisy covariate distributions. These simulations cover a wide range of causal settings such as different ratio of causal effect to background responses, and functional links between covariates, outcome and treatment.

Caussim: Extensive simulation settings

Data Generation. We use simulated data, on which the ground-truth causal effect is known. Going beyond prior empirical studies of causal model selection [50, 51], we use many generative processes, which is needed to reach general conclusions (Appendix A.7).

We generate the response functions using random bases extension, a common method in biostatistics, *e.g.*functional regression

1 Scripts for the simulations and the selection procedure are available at https://github.com/soda-inria/caussim.

with splines [61, 62]. By allowing the function to vary at specific knots, we control the complexity of the non-linear outcome models. We use random approximation of Radial Basis Function (RBF) kernels [63] to generate the outcome and treatment functions. RBF use the same process as polynomial splines but replace polynomial by Gaussian kernels. Unlike polynomial, Gaussian kernels have decreasing influences in the input space. This avoids unrealistic divergences of the functions at the ends of the feature space. We generate 1 000 datasets based on these functions, with random overlap parameters. Example shown in Figure 13 and details in A.7.

Family of candidate estimators. We test model selection across different candidate estimators that approximate imperfectly the data-generating process. To build such estimators, we first use ~~a~~ an RBF expansion similar to that used for data generation. We choose two random knots and transform the raw data features with a Gaussian kernel. This step is referred as the featurization. Then, we fit a linear regression on these transformed features. We consider two ways of combining these steps for outcome model; we use common nomenclature [41, 64] to refer to these different meta-learners that differ on how they model, jointly or not, the treated and the non treated:

- SLearner: A single learner for both ~~population~~populations, taking the treatment as a supplementary covariate.
- SftLearner: A single set of basis functions is sampled at random for both populations, leading to a given feature space used to model both the treated and controls, then two separate different regressors are fitted on this shared representation.
- TLearner: Two completely different learners for each population, hence separate feature representations and regressors.

For the regression step, we fit a Ridge regression on the transformed features with 6 different choices of the regularization parameter $\lambda \in [10^{-3}, 10^{-2}, 10^{-1}, 1, 10^{1}, 10^{2}]$, coupled with a TLearner or a SftLearner. We sample 10 different random basis for learning and featurization yielding a family $\mathcal{F}$ of 120 candidate estimators.

Semi-simulated datasets

Datasets. We also use three semi-simulated data adding a known synthetic causal effect to real –non synthetic– healthcare covariate. ACIC 2016 [45] is based on the Collaborative Perinatal Project [65], a RCT studying infants' developmental disorders containing 4,802 indivduals and 55 features. We used 770 dataset instances: 10 random seeds for each of the 77 simulated settings for the treatment and outcomes. ACIC 2018 [66] simulated treatment and outcomes for the Linked Births and Infant Deaths Database (LBIDD) [67] with $D = 177$ covariates. We used all 432 datasets of size $N = 5\,000$. Twins [68] is an augmentation of real data on twin births and mortality rates [69]. There are $N = 11\,984$ samples, and $D = 50$ covariates for which we simulated 1,000 different treatment allocations. Appendix A.7 gives datasets details.

Family of candidate estimators. For these three datasets, the family of candidate estimators are gradient boosting trees for both the response surfaces and the treatment [2] with S-learner, learning rate in $\{0.01, 0.1, 1\}$, and maximum number of leaf nodes in $\{25, 27, 30, 32, 35, 40\}$ resulting in a family of size 18.

Nuisance estimators. Drawing from the TMLE literature that uses combination of flexible machine learning methods [37], we model the nuisances $\check{e}$ (respectivley $\check{m}$) with a meta-learner: a stacked estimator of ridge and boosting classifiers (respectively regressions) (hyperparameter selection in Appendix A.7).

Measuring overlap between treated and non treated

Good overlap between treated and control population is crucial for causal inference (Assumption 3). We introduce the Normalized Total Variation (NTV), a divergence based on the propensity score summarizing the overlap between both populations (Appendix A.6).

Results: factors driving good model selection

The R-risk is the best metric on average. Figure 3 shows the agreement between the ideal ranking of outcome models given the oracle τ-risk and the different feasible causal metrics. We measure this agreement with relative[3] Kendall tau κ (eq. 20) [70]. Given the importance of overlap in how well metrics approximate the oracle τ-risk, we separate strong and weak overlap.

Among all metrics, the classical mean squared error (ie. factual μ-risk) is worse and reweighting it with propensity score (μ-risk$_{IPW}$) does not bring much improvements. The R-risk, which includes a model of mean outcome and propensity scores, leads to the best performances. Interestingly, the U-risk, which uses the same nuisances, deteriorates in weak overlap, probably due to variance inflation when dividing by extreme propensity scores.

Beyond rankings, the differences in terms of absolute ability to select the best model are large: The R-risk selects a model with a τ-risk only 1% higher than the best possible candidate for strong overlap on Caussim, but selecting with the μ-risk or μ-risk$_{IPW}$ –as per machine-learning practice– leads to 10% excess risk and using τ-risk$_{IPW}$ –as in some causal-inference methods [71, 72]– leads to 100% excess risk (Figure 16). Across datasets, the R-risk consistently decreases the risk compared to the μ-risk: from 0.1% to 1% on ACIC2016, 1% from to 20% on ACIC2018, and 0.05% from to 1% on Twins.

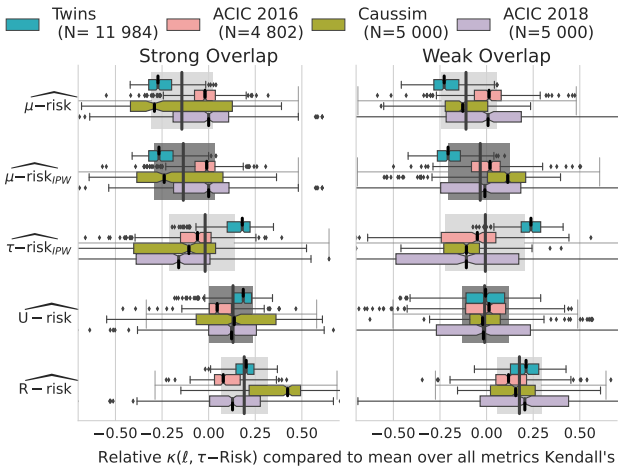


Figure 3. The R-risk is the best metric: Relative Kendall's τ agreement with τ-risk. Strong and Weak overlap correspond to the first and last tertiles of the overlap distribution measured with Normalized Total Variation eq. 17. A.7 presents the same results by adding semi-oracle risks in Figure 14, measured with absolute Kendall's in Figure 15 and with τ−risk gains in Figure 16. Table 4 gives median and IQR of the relative Kendall.

2 Scikit-learn regressor, HistGradientBoostingRegressor, and classifier, HistGradientBoostingClassifier.

3 To remove the variance across datasets (some datasets lead to easier model selection than others), we report values for one metric relative to the mean of all metrics for a given dataset instance: Relative $\kappa(\ell, \tau\text{−risk}) = \kappa(\ell, \tau\text{−risk}) - mean_{\ell}\,(\kappa(\ell, \tau\text{−risk}))$

Model selection is harder for low population overlap. Model selection for causal inference becomes more and more difficult with increasingly different treated and control populations (Figure 4). The absolute Kendall's coefficient correlation with τ-risk drops from 0.9 (excellent agreement with oracle selection) to 0.6 on both Caussim and ACIC 2018 (15).

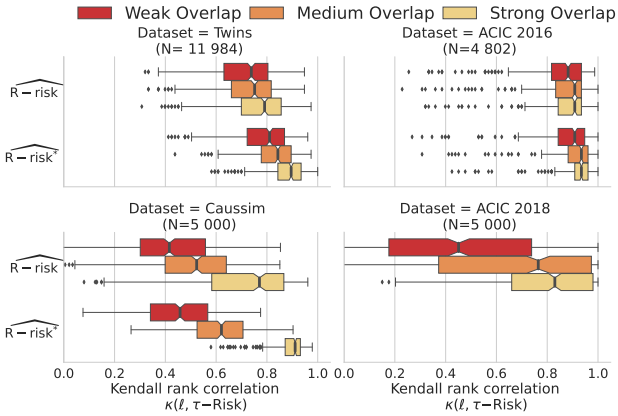


Figure 4. Model selection is harder for low population overlap: Kendall's τ agreement with τ-risk. Strong, medium and Weak overlap are the tertiles of the overlap measured with NTV eq. 17. Supplementary materials presents results for all metrics in Figure 18 in absolute Kendall's and continuous overlap values in Figure 15.

Nuisances can be estimated on the same data as outcome models. Using the train set $\mathcal{T}$ both to fit the candidate estimator and the nuisance estimates is a form of double dipping which can lead errors in nuisances correlated to that of outcome models [42]. In theory, these correlations can bias model selection and, strictly speaking, push to split out a third separated ~~data set~~ dataset –a "nuisance set"– to fit the nuisance models. The drawback is that it depletes the data available for model estimation and selection. However, Figure 5 shows no substantial difference between a procedure with a separated nuisance set and the simpler shared nuisance-candidate set procedure.

Empirically, the best split is 90 %/10 %: using 90 % of the data to estimate both the nuisances and candidates, then computing the risks on the remaining test set for model selection (experiments in Appendix A.8).

Stacked models are good overall estimators of nuisances. Stacked nuisances estimators (boosting and linear) lead to feasible metrics with close performances to the oracles ones: the corresponding estimators recover well-enough the true nuisances. One may wonder if simpler models for the nuisance could be useful, in particular in data-poor settings or when the true models are linear. Figure 6 compares causal model selection estimating nuisances with stacked estimators or linear model. It comprises the Twins data, where the

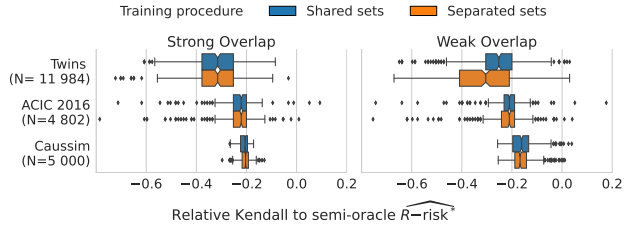


Figure 5. Nuisances can be estimated on the same data as outcome models: Results for the R-risk are similar between the shared nuisances/candidate set and the separated nuisances set procedures. Figure 17 details results for all metrics.

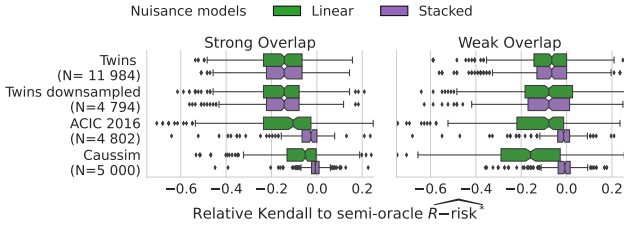


Figure 6. Stacked models are good overall estimators of the nuisances: Results are shown only for the R-risk; Figure 19 details every metrics. For Twins, where the true propensity model is linear, stacked and linear estimations of the nuisances performs equivalently, even for a downsampled version (N=4,794).

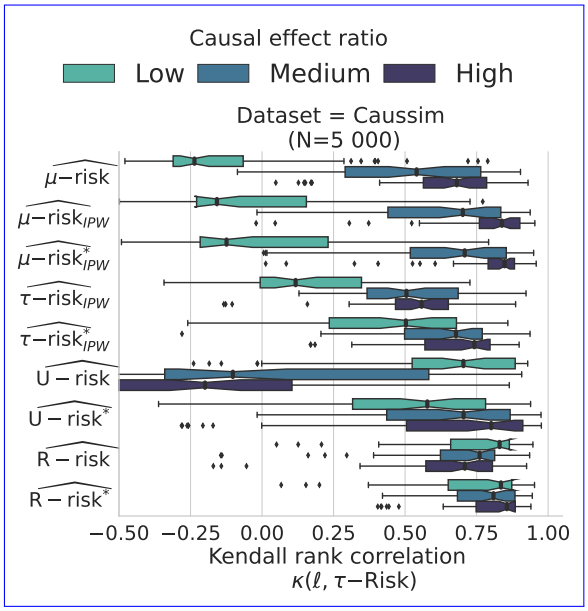


Figure 7. R-risk is robust to a wide range of effect ratio: Kendall's τ agreement with τ-risk. Strong, medium and Weak Causal effect ratio are the tertiles of the absolute ratio causal effect to baseline response, Δ_μ: Low $[0.005; 0.4[$, Medium $[0.44; 9[$, High $[9; 9675[$. Appendix A.9 details this simulation.

true propensity model is linear, and a downsampled version of this data, to study a situation favorable to linear models. In these settings, stacked and linear estimations of the nuisances performs equivalently. Detailed analysis (Figure 20) confirms that using adaptive models –as built by stacking linear models and gradient-boosted trees– suffices to estimate nuisance.

R-risk is robust to a wide range of effect ratio values. Beyond overlap, we study for caussim simulations, the effect on model selection of different causal effect ratio to baseline. We vary the empirical mean absolute difference between the causal effect and the baseline, $\Delta_\mu = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\mu_1(x_i)-\mu_0(x_i)}{\mu_0(x_i)}\right|$, covering a ratio range from 0.05 to 1000 (median = 1.9, full distribution in Appendix A.9). Figure 7 shows that for high values of the ratio, $R-$risk is outperformed by the μ-risk$_{IPW}$ and the τ-risk$_{IPW}$. However, on average, the $R-$risk is still the better risk.

Discussion and conclusion

Nuisance models: more gain than pain. Predictive models are increasingly used to reason about treatment effects, for instance in precision medicine to drive individualized decision. Our results highlight that they should be selected, validated, and tuned using different procedures and error measures than those classically used to assess

prediction. Rather, selecting the best outcome model according to the R-risk (eq. Definition 5) leads to more valid causal estimates on average. Estimating the R-risk requires a more complex procedure than standard cross-validation used *e.g.*in machine learning: it involves fitting nuisance models necessary for model evaluation. Our results show that these can be learned on the same set of data as the outcome models evaluated. The nuisance models must be well estimated (Figure 6). Our results show that using for nuisance models a flexible stacking-based family of estimator suffices for good model selection. To select propensity score models, we used the Brier score, minimized by the true individual probability. An easy mistake is to use calibration errors popular in machine learning [73, 74, 75, 76] as these select not for the individual posterior probability but for an aggregate error rate [77].

More R-risk to select models driving decisions. Increasingly complex prediction models integrating richer medical data have flourished because their predictions can be easily demonstrated and validated on left-out data. But using them to underpin a decision on whether to treat or not requires more careful validation, using a metric accounting for the putative intervention, the R-risk. ~~The~~ On average, the R-risk brings a sizeable benefit to select the most adequate model, even when model development is based on treated and untreated population with little differences, as in RCTs. To facilitate better model selection, we provide Python code[4]. This model-selection procedure puts no constraints on the models used to build predictive models: it opens the door to evaluating a wide range of models, from gradient boosting to convolutional ~~neutral~~neural network, or language models.

Limitations. On average, the R-risk outperforms other risks, but it is not optimal for every situations. If one possesses some prior knowledge on the data generation process, another risk might be more adapted. This is typically the case for big causal effect ratio to baseline and low overlap where the τ-risk$_{IPW}$ should be better.

Availability of source code and requirements

Lists the following:

- Project name: Caussim
- Project home page: `https://github.com/soda-inria/caussim`
- Operating system(s): Platform independent
- Programming language: Python
- License: BSD 3-Clause License

Declaration

Competing interests

No competing interest is declared.

Author contributions statement

M.D. conceived and conducted the experiments, M.D. and G.V. analyzed the results. M.D. and G.V. wrote and reviewed the manuscript.

Acknowledgments

We acknowledge fruitful discussions with Bénédicte Colnet.

4 `https://github.com/soda-inria/causal_model_selection`

A Additional files

A.1 Variability of ATE estimation on ACIC 2016

Figure 8 shows ATE estimations for six different models used in g-computation estimators on the 76 configurations of the ACIC 2016 dataset. Outcome models are fitted on half of the data and inference is done on the other half –ie. train/test with a split ratio of 0.5. For each configuration, and each model, this train test split was repeated ten times, yielding non parametric variance estimates [78]. Figure 8 shows large variations obtained across different outcome estimators on semi-synthetic datasets [45]. Flexible models such as random forests are doing well in most settings except when treated and untreated populations differ noticeably, in which case a linear model (ridge) is to be preferred. However random forests with different hyper-parameters (max depth= 2) yield poor estimates. A simple rule of thumb such as preferring flexible models does not work in general; model selection is needed.

Outcome models are implemented with scikit-learn [79] and the following hyper-parameters:

Outcome Model	Hyper-parameters grid
Random Forests	Max depth: [2, 10]
Ridge regression without treatment interaction	Ridge regularization: [0.1]
Ridge regression with treatment interaction	Ridge regularization: [0.1]

Table 2. Hyper-parameters grid used for ACIC 2016 ATE variability

A.2 Prior work : model selection for outcome modeling (g-computation)

A natural way to select a predictive model for causal inference would be an error measure between a causal quantity such as the CATE and models' estimate. But such error is not a "feasible" risk: it cannot be computed solely from observed data and requires oracle knowledge.

Simulation studies of causal model selection. Using eight simulations setups from [38], where the oracle CATE is known, [50] compare four causal risks, concluding that for CATE estimation the best model-selection risk is the so-called R-risk [42] –def. 5, below. Their empirical results are clear for randomized treatment allocation but less convincing for observational settings where both simple ~~Mean Squared Error –MSE,~~ MSE – μ-risk(f) def. 1– and reweighted MSE – μ-risk$_{IPW}$ def. 2– appear to perform better than R-risk on half of the simulations. Another work [51] studied empirically both MSE and reweighted MSE risks on the semi-synthetic ACIC 2016 datasets [45], but did not include the R-risk. We complete these prior empirical work by studying a wider variety of data generative processes and varying the

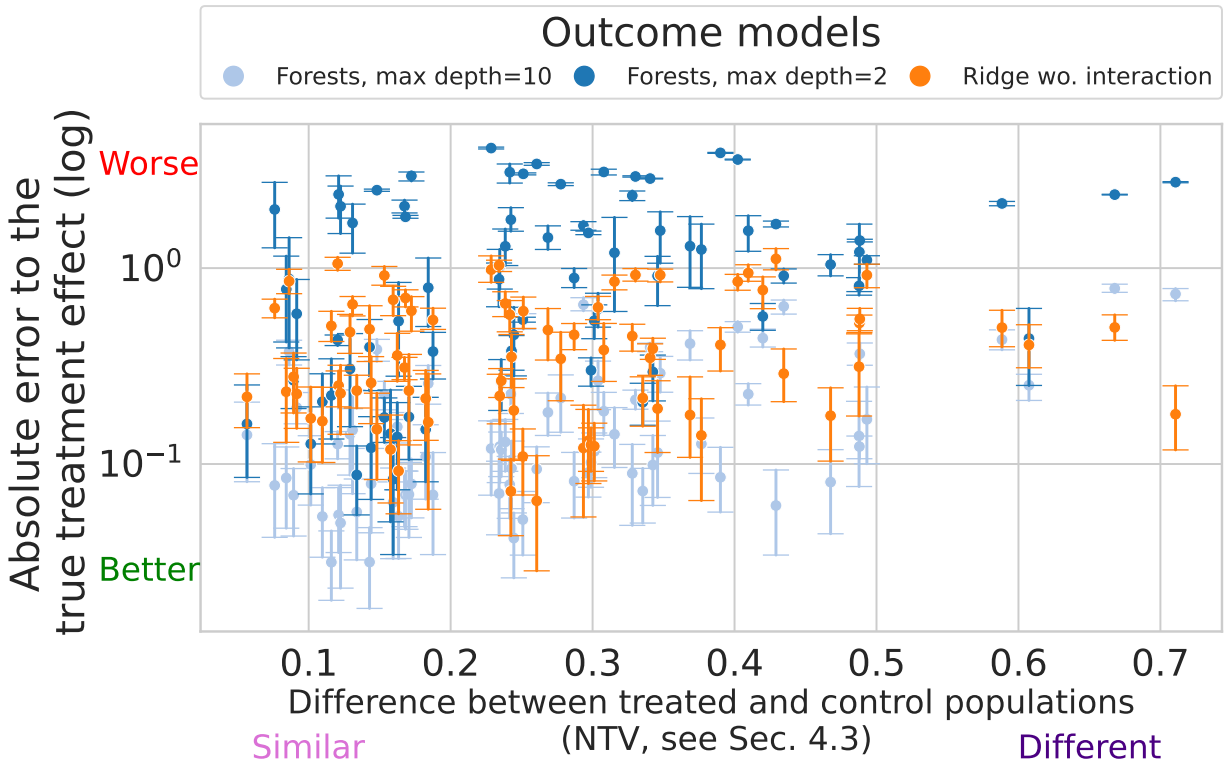


Figure 8. Different outcome models lead to different estimation errors on the Average Treatment Effects, on 77 classic simulations with known true causal effect [45]. The different models are ridge regression and random forests with different hyper-parameters (details A.1). The different configurations are plotted as a function of increasing difference between treated and untreated population –see subsection . There is no systematic best performer; data-driven model selection is important.

influence of overlap, an important parameter of the data generation process which makes a given causal metric appropriate [80]. We also study how to best adapt cross-validation procedures to causal metrics which themselves come with models to estimate.

Theoretical studies of causal model selection. Several theoretical works have proposed causal model selection procedures that are *consistent*: select the best model in a family given asymptotically large data. These work rely on introducing a CATE estimator in the testing procedure: matching [81], an IPW estimate [72], a doubly robust estimator [82], or debiasing the error with influence functions [51]. However, for theoretical guarantees to hold, the test-set correction needs to converge to the oracle: it needs to be flexible enough –well-posed– and asymptotic data. From a practical perspective, meeting such requirements implies having a good CATE estimate, thus having solved the original problem of causal model selection.

Statistical guarantees on causal estimation procedures. Much work in causal inference has focused on procedures that guarantee asymptotically consistent estimators, such as Targeted Machine Learning Estimation (TMLE) [36, 37] or Double Machine Learning [43]. Here also, theories require asymptotic regimes and models to be *well-specified*.

By contrast, without assuming that estimators are well specified, there exists an upper bound on the oracle error to the CATE (τ-risk) that involves the error on the outcome and the similarity of the distributions of treated and control patients [83]. However, they use this upper bound for model optimization, and do not give insights on model selection. In addition, for hyperparameter selection, they rely on a plugin estimate of the τ-risk built with counterfactual nearest neighbors, which has been shown ineffective [50]. An interesting direction is taken in [84] where the authors derive convergence rates for orthogonal losses such as the R-loss, or the DR-Loss without assuming well-specification of the model for target parameter.

A.3 Causal assumptions

We assume the following four assumptions, referred as strong ignorability and necessary to assure identifiability of the causal estimands with observational data [56]:

Assumption 1 (Consistency) *The observed outcome is the potential outcome of the assigned treatment:*

$$Y = A\,Y(1) + (1 - A)\,Y(0)$$

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 [24].

Assumption 2 (Unconfoundedness)

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

This condition –also called ignorability– is equivalent to the conditional independence on $e(X)$ [28]: $\{Y(0), Y(1)\} \perp\!\!\!\perp A|e(X)$.

Assumption 3 (Overlap, also known as Positivity))

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

The treatment is not perfectly predictable. Or with different 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 [80], the choice of covariates X can be viewed as a trade-off between these two central assumptions. A bigger covariates set generally reinforces the ignorability assumption. In the contrary, overlap can be weakened by large $\mathcal{X}$ because of the potential inclusion of instruments: variables only linked to the treatment which could lead to arbitrarily small propensity scores.

~~The observed outcome is the potential outcome of the assigned treatment:~~

$$Y = A\,Y(1) + (1 - A)\,Y(0)$$

~~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 [24].~~

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 $\mathcal{D}^*$, also known as the "no covariate shift" assumption [85].*

A.4 Definitions of feasible risks

Definition 1 (Factual μ-risk) *[60] This is the usual Mean Squared Error on the target y. It is what is typically meant by "generalization error" in supervised learning:*

$$\mu\text{-}risk(f) = \mathbb{E}\left[(Y - f(X;A))^2\right]$$

Definition 2 (μ-risk$^{\star}_{IPW}$) *[58] Let the inverse propensity weighting function $w(x,a) = \frac{a}{e(x)} + \frac{1-a}{1-e(x)}$, we define the semi-oracle Inverse Propensity Weighting risk,*

$$\mu\text{-}risk^{\star}_{IPW}(f) = \mathbb{E}\left[\left(\frac{A}{e(X)} + \frac{1-A}{1-e(X)}\right)(Y - f(X;A))^2\right]$$

Definition 3 (τ-risk$^{\star}_{IPW}$) *[39] The CATE $\tau(x)$ can be estimated with a regression against inverse propensity weighted outcomes [71, 72, 39], the τ-risk$_{IPW}$.*

$$\tau\text{-}risk^{\star}_{IPW}(f) = \mathbb{E}\left[\left(Y\frac{A - e(X)}{e(X)(1-e(X))} - \tau_f(X)\right)^2\right]$$

Definition 4 (U-risk*) *[41, 42] Based on the Robinson decomposition –eq. 8, the U-learner uses the $A - e(X)$ term in the denominator. The derived risk is:*

$$U\text{-}risk^{\star}(f) = \mathbb{E}\left[\left(\frac{Y - m(X)}{A - e(X)} - \tau_f(X)\right)^2\right]$$

Note that extreme propensity weights in the denominator term might inflate errors in the numerator due to imperfect estimation of the mean outcome m.

Definition 5 (R-risk*) *[42, 50] The R-risk also uses two nuisance m and e:*

$$R\text{-}risk^{\star}(f) = \mathbb{E}\left[\left((Y - m(X)) - (A - e(X))\tau_f(X)\right)^2\right]$$

It is also based on the Robinson decomposition –eq. 8.

A.5 Proofs: Links between feasible and oracle risks

Reformulation of the R-risk as reweighted τ-risk
Proposition 1 (R-risk as reweighted τ-risk) ***Proof 1*** *We consider the R-decomposition: [57],*

$$y(a) = m(x) + (a - e(x))\tau(x) + \varepsilon(x;a) \tag{12}$$

Where $\mathbb{E}[\varepsilon(X;A)|X,A] = 0$ We can use it as plug in the R-risk formula:

$$\begin{aligned}
R\text{-}risk(f) &= \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{A}} [(y - m(x)) - (a - e(x))\tau_f(x)]^2 p(y;x;a)dydxda \\
&= \int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{A}} \left[(a - e(x))\tau(x) + \varepsilon(x;a) - (a - e(x))\tau_f(x)\right]^2 p(y;x;a)dydxda \\
&= \int_{\mathcal{X}\times\mathcal{A}} (a - e(x))^2(\tau(x) - \tau_f(x))^2 p(x;a)dxda \\
&+ 2\int_{\mathcal{Y}\times\mathcal{X}\times\mathcal{A}} (a - e(x))(\tau(x) - \tau_f(x)) \int_{\mathcal{Y}} \varepsilon(x;a)p(y \mid x;a)dyp(x;a)dxda \\
&+ \int_{\mathcal{X}\times\mathcal{A}} \int_{\mathcal{Y}} \varepsilon^2(x;a)p(y \mid x;a)dyp(x;a)dxda
\end{aligned}$$

The first term can be decomposed on control and treated populations to force $e(x)$ to appear:

$$\begin{aligned}
&\int_{\mathcal{X}} (\tau(x) - \tau_f(x))^2 \left[e(x)^2 p(x;0) + (1 - e(x))^2 p(x;1)\right] dx \\
&= \int_{\mathcal{X}} (\tau(x) - \tau_f(x))^2 \left[e(x)^2(1 - e(x))p(x) + (1 - e(x))^2 e(x)p(x)\right] dx \\
&= \int_{\mathcal{X}} (\tau(x) - \tau_f(x))^2(1 - e(x))e(x)[1 - e(x) + e(x)]p(x)dx \\
&= \int_{\mathcal{X}} (\tau(x) - \tau_f(x))^2(1 - e(x))e(x)p(x)dx.
\end{aligned}$$

The second term is null since, $\mathbb{E}[\varepsilon(x,a)|X,A] = 0$.
The third term corresponds to the modulated residuals : $\tilde{\sigma}^2_B(0) + \tilde{\sigma}^2_B(1)$

Interesting special cases
Randomization special case. If the treatment is randomized as in RCTs, $p(A = 1 \mid X = x) = p(A = 1) = p_A$, thus μ-risk$_{IPW}$ takes a simpler form:

$$\mu\text{-risk}_{IPW} = \mathbb{E}_{(Y,X,A)\sim\mathcal{D}}\left[\left(\frac{A}{p_A} + \frac{1-A}{1-p_A}\right)(Y - f(X;A))^2\right]$$

However, we still can have large differences between τ-risk and μ-risk$_{IPW}$ coming from heterogeneous errors between populations as shown experimentally in [50] and our results below.

Concerning the R-risk, replacing $e(x)$ by its randomized value p_A in Proposition 1 yields the oracle τ-risk up to multiplicative and additive constants:

$$R\text{-risk} = p_A\,(1-p_A)\,\tau\text{-risk} \;+\; (1-p_A)\,\sigma_B^2(0) \;+\; p_A\,\sigma_B^2(1)$$

Thus, selecting estimators with R-risk* in randomized setting controls the τ-risk. This explains the strong performances of R-risk in randomized setups [50] and is a strong argument to use it to estimate heterogeneity in RCTs.

Oracle Bayes predictor. If we have access to the oracle Bayes predictor for the outcome ie. $f(x,a) = \mu(x,a)$, then all risks are equivalent up to the residual variance:

$$\tau\text{-risk}(\mu) = \mathbb{E}_{X\sim p(X)}[(\tau(X) - \tau_\mu(X))^2] = 0 \tag{13}$$

$$\begin{aligned}\mu\text{-risk}(\mu) &= \mathbb{E}_{(Y,X,A)\sim p(Y,X,A)}[\,(Y - \mu_A(X))^2\,] \\ &= \int_{\mathcal{X},\mathcal{A}} \varepsilon(x,a)^2 p(a \mid x)\, p(x)\, dx\, da \le \sigma_B^2(0) + \sigma_B^2(1)\end{aligned} \tag{14}$$

$$\mu\text{-risk}_{IPW}(\mu) = \sigma_B^2(0) + \sigma_B^2(1) \quad \text{from Lemma ??} \tag{15}$$

$$R\text{-risk}(\mu) = \tilde{\sigma}_B^2(0) + \tilde{\sigma}_B^2(1) \le \sigma_B^2(0) + \sigma_B^2(1)$$

from Proposition 1 (16)

Thus, differences between causal risks only matter in finite sample regimes. Universally consistent learners converge to the Bayes risk in asymptotic regimes, making all model selection risks equivalent. In practice however, choices must be made in non-asymptotic regimes.

A.6 Measuring overlap

Motivation of the Normalized Total Variation. Overlap is often assessed by comparing visually population distributions as in Figure 1 or computing standardized difference on each feature [86, 29]. While these methods are useful to decide if positivity holds, they do not yield a single measure. Rather, we compute the divergence between the population covariate distributions $\mathbb{P}(X|A=0)$ and $\mathbb{P}(X|A=1)$ [80, 83]. Computing overlap when working only on samples of the observed distribution, outside of simulation, requires a sophisticated estimator of discrepancy between distributions, as two data points never have the same exact set of features. Maximum Mean Discrepancy [87] is typically used in the context of causal inference [60, 83]. However it needs a kernel, typically Gaussian, to extrapolate across neighboring observations. We prefer avoiding the need to specify such a kernel, as it must be adapted to the data which is tricky with categorical or non-Gaussian features, a common situation for medical data.

For simulated and some semi-simulated data, we have access to the probability of treatment for each data point, which sample both densities in the same data point. Thus, we can directly use distribution discrepancy measures and rely on the Normalized Total Variation (NTV) distance to measure the overlap between the treated and control propensities. This is the empirical measure of the total variation distance [88] between the distributions, $TV(\mathbb{P}(X|A=1), \mathbb{P}(X|A=0))$. As we have both distribution sampled on the same points, we can rewrite it a sole function of the propensity score, a low dimensional score more tractable than the full distribution $\mathbb{P}(X|A)$:

$$\widehat{NTV}(e, 1-e) = \frac{1}{2N}\sum_{i=1}^{N} \left|\frac{e(x_i)}{p_A} - \frac{1-e(x_i)}{1-p_A}\right| \tag{17}$$

Formally, we can rewrite NTV as the Total Variation distance between the two population distributions. For a population $O = (Y(A), X, A) \sim \mathcal{D}$:

$$\begin{aligned}NTV(O) &= \frac{1}{2N}\sum_{i=1}^{N} \left|\frac{e(x_i)}{p_A} - \frac{1-e(x_i)}{1-p_A}\right| \\ &= \frac{1}{2N}\sum_{i=1}^{N} \left|\frac{P(A=1|X=x_i)}{p_A} - \frac{P(A=0|X=x_i)}{1-p_A}\right|\end{aligned}$$

Thus NTV approximates the following quantity in expectation over the data distribution $\mathcal{D}$:

$$
\begin{aligned}
NTV(\mathcal{D}) &= \int_{\mathcal{X}} \left| \frac{p(A=1|X=x)}{p_A} - \frac{p(A=0|X=x)}{1-p_A} \right| p(x)dx \\
&= \int_{\mathcal{X}} \left| \frac{p(A=1, X=x)}{p_A} - \frac{p(A=0, X=x)}{1-p_A} \right| dx \\
&= \int_{\mathcal{X}} \left| p(X=x|A=1) - p(X=x|A=0) \right| dx
\end{aligned}
$$

For countable sets, this expression corresponds to the Total Variation distance between treated and control populations covariate distributions : $TV(p_0(x), p_1(x))$.

Measuring overlap without the oracle propensity scores:. For ACIC 2018, or for non-simulated data, the true propensity scores are not known. To measure overlap, we rely on flexible estimations of the Normalized Total Variation, using gradient boosting trees to approximate the propensity score. Empirical arguments for this plug-in approach is given in Figure 9.

Empirical arguments. We show empirically that NTV is an appropriate measure of overlap by :

- Comparing the NTV distance with the MMD for Caussim which is gaussian distributed in Figure 11,
- Verifying that setups with penalized overlap from ACIC 2016 have a higher total variation distance than unpenalized setups in Figure 10.
- Verifying that the Inverse Propensity Weights extrema (the inverse of the ν overlap constant appearing in the overlap Assumption 3) positevely correlates with NTV for Caussim, ACIC 2016 and Twins in Figure 12. Even if the same value of the maximum IPW could lead to different values of NTV, we expect both measures to be correlated : the higher the extrem propensity weights, the higher the NTV.

Estimating NTV in practice. Finally, we verify that approximating the NTV distance with a learned plug-in estimates of e(x) is reasonnable. We used either a logistic regression or a gradient boosting classifier to learn the propensity models for the three datasets where we have access to the ground truth propensity scores: Caussim, Twins and ACIC 2016. We respectively sampled 1000, 1000 and 770 instances of these datasets with different seeds and overlap settings. We first run a hyperparameter search with cross-validation on the train set, then select the best estimator. We refit on the train set this estimator with or without calibration by cross validation and finally estimate the normalized TV with the obtained model. This training procedure reflects the one described in Algorithm 1 where nuisance models are fitted only on the train set.

The hyper parameters are : learning rate $\in [1e-3, 1e-2, 1e-1, 1]$, minimum samples leaf $\in [2, 10, 50, 100, 200]$ for boosting and L2 regularization $\in [1e-3, 1e-2, 1e-1, 1]$ for logistic regression.

Results in Figure 9 comparing bias to the true normalized Total Variation of each dataset instances versus growing true NTV indicate that calibration of the propensity model is crucial to recover a good approximation of the NTV.

A.7 Experiments

Details on the data generation process

We use Gaussian-distributed covariates and random basis expansion based on Radial Basis Function kernels. A random basis of RBF kernel enables modeling non-linear and complex relationships between covariates in a similar way to the well known spline expansion. The estimators of the response function are learned with a linear model on another random basis (which can be seen as a stochastic approximation of the full data kernel [63]). We carefully control the amount of overlap between treated and control populations, a crucial assumption for causal inference. Algorithm 2 describes the generation process for one simulation. Figure 13 illustrates 2D examples of the simulation.

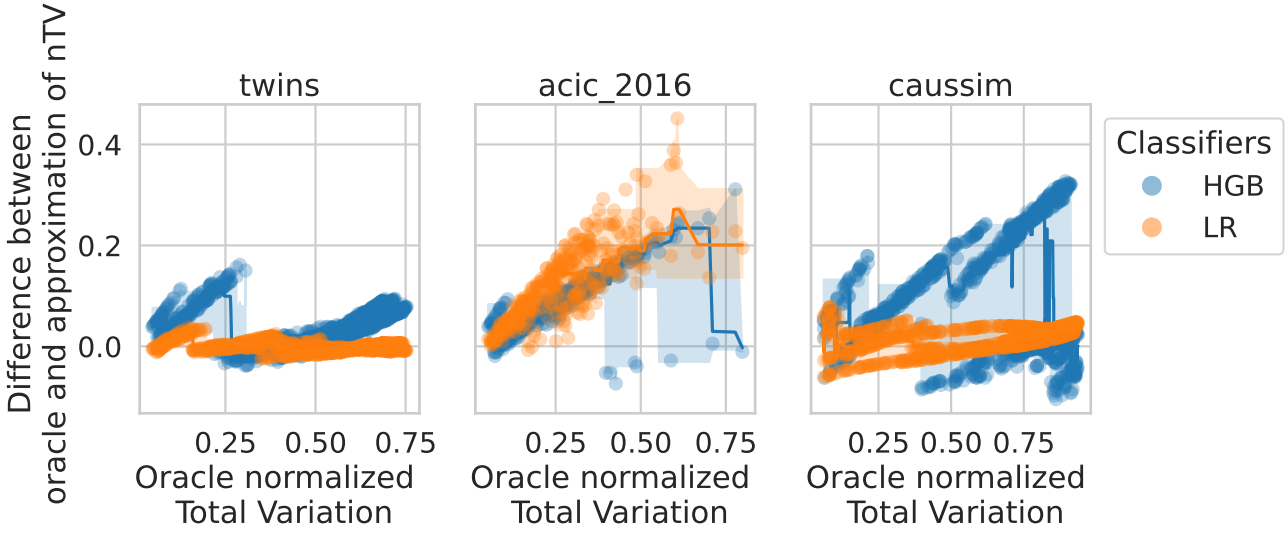
- The raw features for both populations are drawn from a mixture of Gaussians: $\mathbb{P}(X) = p_A\mathbb{P}(X|A=1) + (1-p_A)\mathbb{P}(X|A=0)$ where $\mathbb{P}(x|A=a)$ is a rotated Gaussian:

$$\mathbb{P}(x|A=a) = W \cdot \mathcal{N}\left(\begin{bmatrix} (1-2a)\theta \\ 0 \end{bmatrix}; \begin{bmatrix} \sigma_0 & 0 \\ 0 & \sigma_1 \end{bmatrix} \right) \tag{18}$$

  with θ a parameter controlling overlap (bigger yields poorer overlap), W a random rotation matrix and $\sigma_0^2 = 2; \sigma_1^2 = 5$.
  This generation process allows to analytically compute the oracle propensity scores $e(x)$, to simply control for overlap with the parameter θ, the distance between the two Gaussian main axes and to visualize response surfaces.
- A basis expansion of the raw features increases the problem dimension. Using Radial Basis Function (RBF) Nystroem transformation [5], we expand the raw features into a transformed space. The basis expansion samples randomly a small number of representers in the raw data. Then, it computes an approximation of the full N-dimensional kernel with these basis components, yielding the transformed features $z(x)$. The number of basis functions –*ie. knots*–, controls the complexity of the ground-truth response surfaces and treatment. We first use this process to draw the non-treated response surface μ_0 and the causal effect τ. We then draw the observations from a mixture two Gaussians, for the treated and non treated. We vary the separation between the two Gaussians to control the overlap between treated and non-treated populations, an important parameter for causal inference (related to η in ~~Proposition ??~~Assumption 3). Finally, we generate observed outcomes adding Gaussian noise.
  More formally, we generate the basis following the original data distribution, $[b_1..b_D] \sim \mathbb{P}(x)$, with D=2 in our simulations. Then, we compute an approximation of the full kernel of the data generation process ~~*RBF(x, .) with x ~ ℙ(x)*~~ with these representers: $z(x) =$

5 We use the Sklearn implementation, [79]

(a) Uncalibrated classifiers



(b) Calibrated classifiers

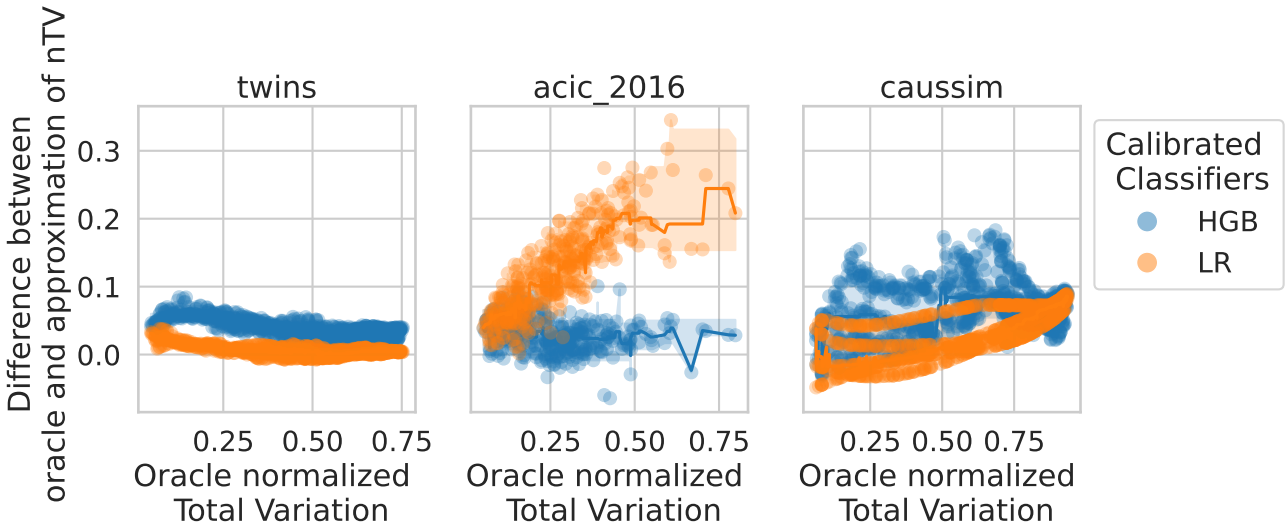


Figure 9. a) Without calibration, estimation of NTV is not trivial even for boosting models. b) Calibrated classifiers are able to recover the true Normalized Total Variation for all datasets where it is available.

Figure 10. NTV recovers well the overlap settings described in the ACIC paper [45]

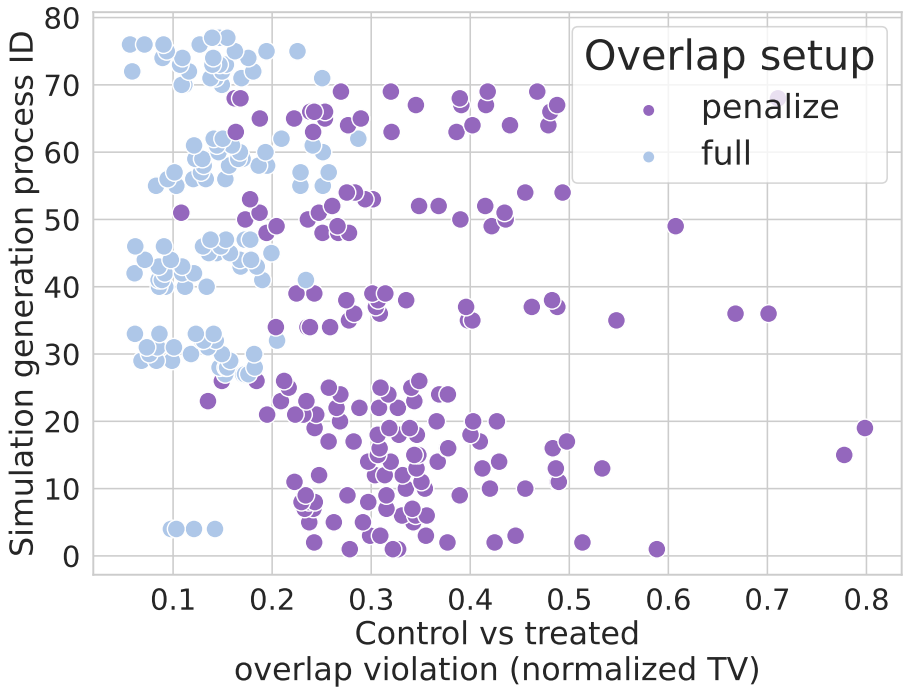
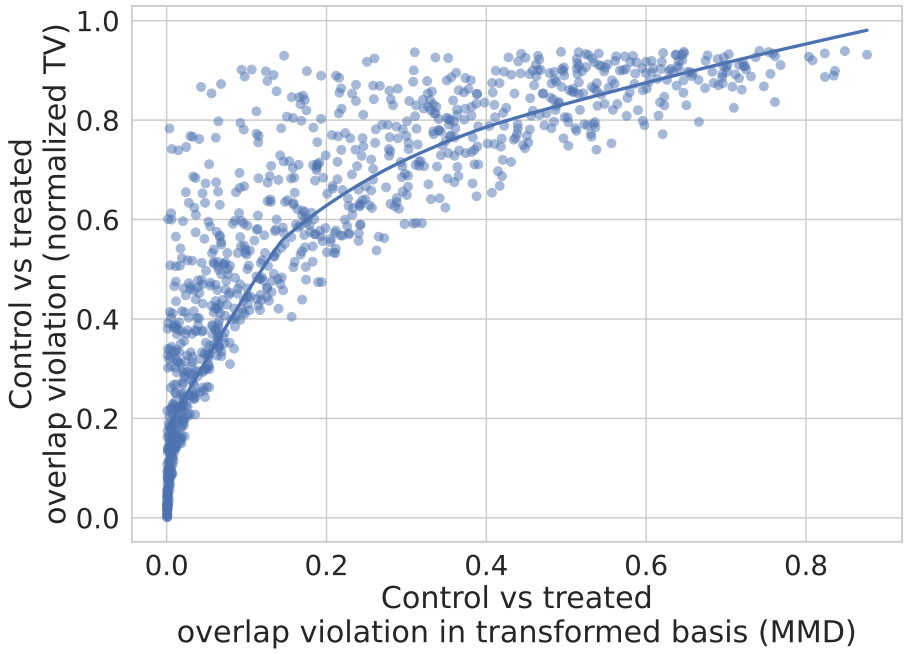


Figure 11. Good correlation between overlap measured as normalized Total Variation and Maximum Mean Discrepancy (200 sampled Caussim datasets)



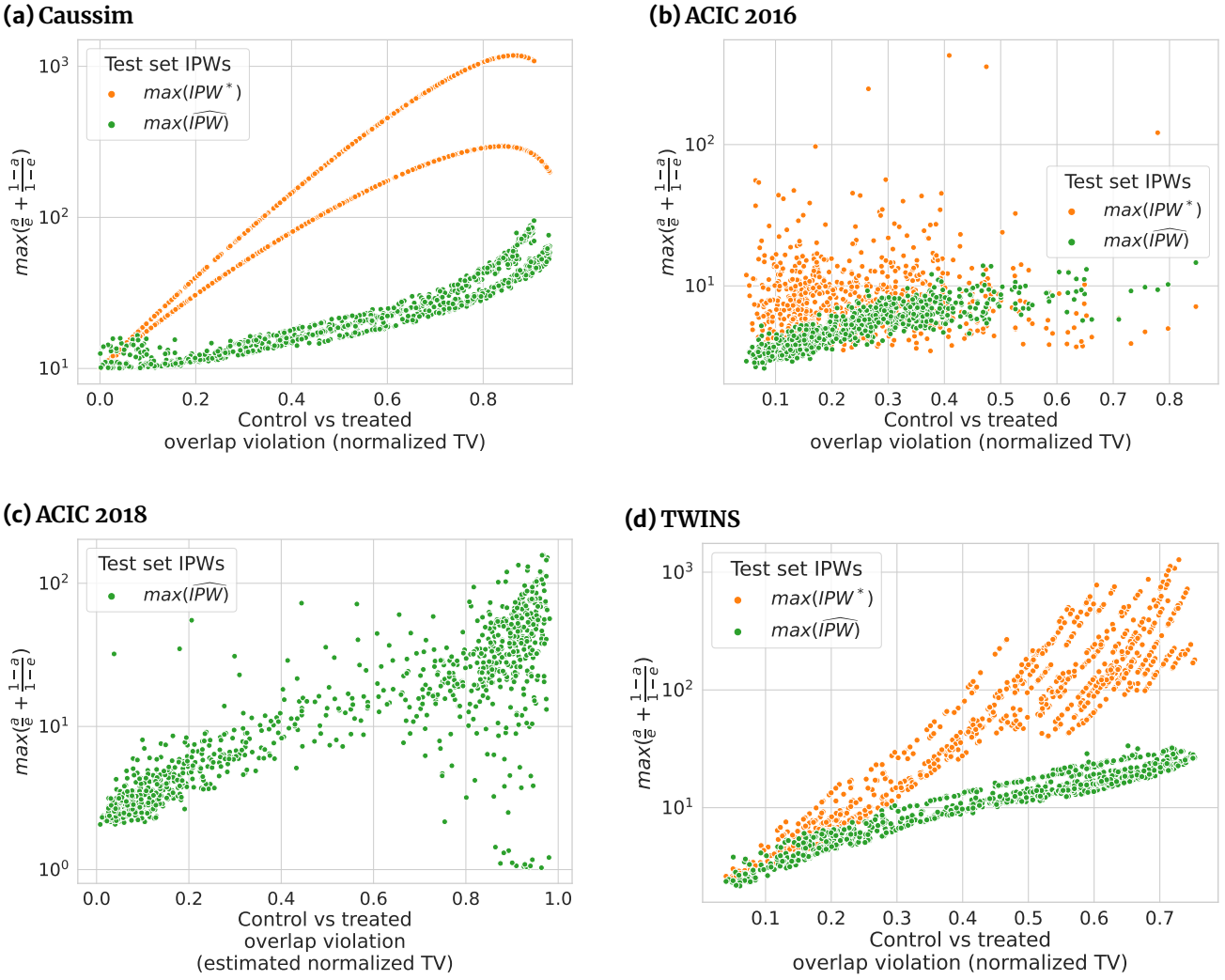


Figure 12. Maximal value of Inverse Propensity Weights increases exponentially with the overlap as measure by Normalized Total Variation.

$[RBF_\gamma(x, b_d)]_{d=1..D} \cdot Z^T \in \mathbb{R}^D$ with RBF_γ being the Gaussian kernel $K(x, y) = exp(-\gamma||x - y||^2)$ and Z the normalization constant of the kernel basis, computed as the root inverse of the basis kernel $Z = [K(b_i, b_j)]^{-1/2}_{i,j \in 1..D}$

- Functions μ_0, τ are distinct linear functions of the transformed features:

$$\mu_0(x) = \left[z(x); 1\right] \cdot \beta_\mu^T$$

$$\tau(x) = \left[z(x); 1\right] \cdot \beta_\tau^T$$

Where β_μ and β_τ are sampled from two normal distributions with mean 0 and unit variances $\mathcal{N}(0, I_{D+1})$.

- Adding a Gaussian noise, $\varepsilon \sim \mathcal{N}(0, \sigma(x; a))$, we construct the potential outcomes: ~~$y(a) = \mu_0(x) + a\,\tau(x) + \varepsilon(x, a)$~~ $y(a) = (1 - \omega)\mu_0(x) + a\,\omega\tau(x) + \varepsilon(x, a)$ with ω the treatment effect relative size with respect to the baseline response.

We generated 1000 instances of this dataset with uniformly random overlap parameters $\theta \in [0, 2.5]$.

Details on the semi-simulated datasets

ACIC 2018 [45]: The initial intervention was a child's birth weight ($A = 1$ if weight $< 2.5kg$), and outcome was the child's IQ after a follow-up period. The study contained $N = 4\,802$ data points with $D = 55$ features (5 binary, 27 count data, and 23 continuous). They simulated 77 different setups varying parameters for treatment and response models, overlap, and interactions between treatment and covariates [6]. We used 10 different seeds for each setup, totaling 770 dataset instances.

ACIC 2018 [66]: Starting from data from the Linked Births and Infant Deaths Database (LBIDD) [67] with $D = 177$ covariates, treatment and outcome models are simulated with complex models to reflect different scenarii. The data do not provide the true propensity scores, so

6 Original R code available at https://github.com/vdorie/aciccomp/tree/master/2016 to generate 77 simulations settings.

Algorithm 2 Data simulation for the simulated dataset caussim

Let the parameters of a given simulation be:

- N the number of samples.
- D the dimension of the basis expansion.
- p_A the proportion of treated individuals.
- θ the overlap parameter.
- $\mathbb{P}(x|A=a) = W \cdot \mathcal{N}\left(\begin{bmatrix}(1-2a)\theta \\ 0\end{bmatrix}; \begin{bmatrix}\sigma_0 & 0 \\ 0 & \sigma_1\end{bmatrix}\right)$ the rotated gaussian for each population covariate. W is a random rotation matrix and $\sigma_0^2 = 2; \sigma_1^2 = 5$ are scaling parameters.
- $e(x) = \mathbb{P}(A=1|X=x)$ the oracle propensity score, obtained analytically from $\mathbb{P}(x|A=a)$.
- $\mathbb{P}(X) = p_A\mathbb{P}(X|A=1) + (1-p_A)\mathbb{P}(X|A=0)$ the mixture of gaussian for the whole population covariates.
- $b_1, .., b_D \sim \mathbb{P}(X)$, the basis sampled from the gaussian mixture.
- $z(x) = [RBF_\gamma(x, b_d)]_{d=1..D} \cdot Z^T \in \mathbb{R}^D$ the Nystroem expansion. RBF_γ is the Gaussian kernel $K(x,y) = exp(-\gamma||x-y||^2)$ and Z is the normalization constant of the kernel basis, computed as the root inverse of the basis kernel $Z = [K(b_i, b_j)]^{-1/2}_{i,j\in 1..D}$.
- β_μ and β_τ, the linear coefficients on top of the basis expansion. They are sampled from two normal distributions with mean 0 and unit variances $\mathcal{N}(0, I_{D+1})$.
- ω the treatment effect relative size with respect to the baseline response.
- $\varepsilon(x,a) \sim \mathcal{N}(0, \sigma_y)$ an exogen gaussian noise with σ_y the scale of the noise.

Generation process for one sample $(x, a, e(x), y(0), y(1), y)$:

1: $x \sim \mathbb{P}(X)$ ▷ Sample from rotated gaussian mixture
2: $z(x)$ ▷ Compute the Nystroem expansion
3: $\mu_0(x) = \left[z(x); 1\right] \cdot \beta_\mu^T$ ▷ Compute the non-treated response surface, ie. the baseline
4: $\tau(x) = \left[z(x); 1\right] \cdot \beta_\tau^T$ ▷ Compute the conditional treatment effect (CATE)
5: $y(a) = (1-\omega)\mu_0(x) + a\,\omega\tau(x) + \varepsilon(x,a) + \varepsilon(x,a)$ ▷ Compute the potential outcomes
6: $a \sim \text{Bernoulli}(e(x))$ ▷ Compute the treatment status
7: $y = a\,y(1) + (1-a)\,y(0)$ ▷ Compute the observed outcome

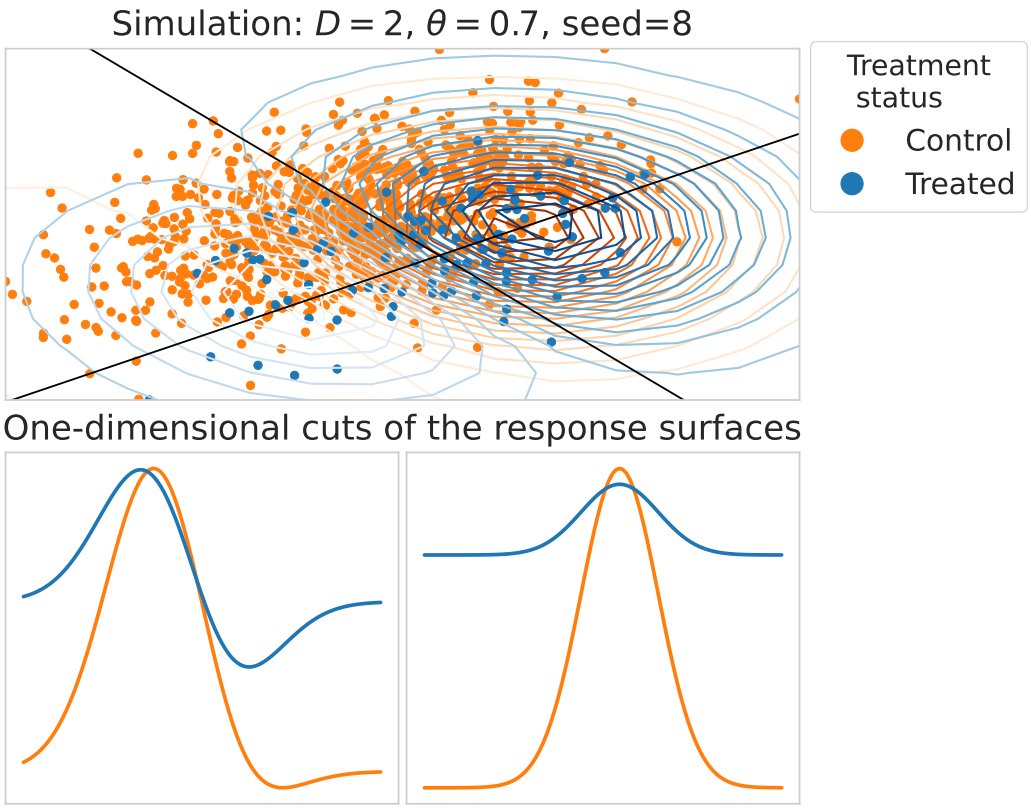


Figure 13. Example of the simulation setup in the input space with two knots –*ie.* basis functions. The top panel shows the observations in feature space, while the bottom panel displays the two response surfaces on a 1D cut along the black lines drawn on the top panel.

we evaluate only feasible metrics, which do not require this nuisance parameter. We used all 432 datasets[7] of size $N = 5\,000$.

7 Using the scaling part of the data, from github.com/IBM-HRL-MLHLS/IBM-Causal-Inference-Benchmarking-Framework

Twins [68]: It is an augmentation of real data on twin births and mortality rates [69]. There are $N = 11\,984$ samples (pairs of twins), and $D = 50$ covariates[8], The outcome is the mortality and the treatment is the weight of the heavier twin at birth. This is a "true" counterfactual dataset [89] in the sense that we have both potential outcomes with each twin. They simulate the treatment with a sigmoid model based on GESTAT10 (number of gestation weeks before birth) and x the 45 other covariates:

$$\mathbf{t}_i \mid \mathbf{x}_i, \mathbf{z}_i \sim \text{Bern}\left(\sigma\left(w_o^\top \mathbf{x} + w_h(\mathbf{z}/10 - 0.1)\right)\right) \quad \text{with } w_o \sim \mathcal{N}(0, 0.1 \cdot I),\ w_h \sim \mathcal{N}(5, 0.1) \tag{19}$$

We add a non-constant slope in the sigmoid to control the overlap between treated and control populations. We sampled uniformly 1 000 different overlap parameters between 0 and 2.5, totaling 1 000 dataset instances. Unlike the previous datasets, only the overlap varies for these instances. The response surfaces are set by the original outcomes.

Model selection procedures

Nuisances estimation. The nuisances are estimated with a stacked regressor inspired by the Super Learner framework, [90]). We select hyper-parameters with randomized search on a validation set $\mathcal{V}$ and keep them fix for model selection The searc grid is detailed in Table 3. All implementations come from scikit-learn [79]. As extreme inverse propensity weights induce high variance, we use clipping [91, 92] to bound $min(\check{e}, 1 - \check{e})$ away from 0 with a fixed $\eta = 10^{-10}$, ensuring strict overlap for numerical stability.

Model	Estimator	Hyper-parameters grid
Outcome, m	StackedRegressor (HistGradientBoostingRegressor, ridge)	ridge regularization: [0.0001, 0.001, 0.01, 0.1, 1, 10, 100] HistGradientBoostingRegressor learning rate: [0.01, 0.1, 1] HistGradientBoostingRegressor max leaf nodes: [10, 20, 30, 50]
Treatment, e	StackedClassifier (HistgradientBoostingClassifier, LogisticRegression)	LogisticRegression C: [0.0001, 0.001, 0.01, 0.1, 1, 10, 100] HistGradientBoostingClassifier learning rate: [0.01, 0.1, 1] HistGradientBoostingClassifier max leaf nodes: [10, 20, 30, 50]

Table 3. Hyper-parameters grid used for nuisance models

Additional Results

Definition of the Kendall's tau, κ. The Kendall's tau is a widely used statistics to measure the rank correlation between two set of observations. It measures the number of concordant pairs minus the discordant pairs normalized by the total number of pairs. It takes values in the $[-1, 1]$ range.

$$\kappa = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{(\text{number of pairs})} \tag{20}$$

Values of relative $\kappa(\ell, \tau\text{–risk})$ compared to the mean over all metrics Kendall's as shown in the boxplots of Figure 3.

Figure 14 - Results measured in relative Kendall's for feasible and semi-oracle risks. Because of extreme propensity scores in the denominator and bayes error residuals in the numerator, the semi-oracle U-risk has poor performances at bad overlap. Estimating these propensity scores in the is feasible U-risk reduces the variance since clipping is performed.

Figure 15 - Results measured in absolute Kendall's.

Figure 16 - Results measured as distance to the oracle tau-risk. To see practical gain in term of τ-risk, we plot the results as the normalized distance between the estimator selected by the oracle τ-risk and the estimator selected by each causal metric.

Then, $\widehat{R\text{-risk}}^*$ is more efficient than all other metrics. The gain are substantial for every datasets.

Figure 17 - Stacked models for the nuisances is more efficient. For each metrics the benefit of using a stacked model of linear and boosting estimators for nuisances compared to a linear model. The evaluation measure is Kendall's tau relative to the oracle R-risk* to have a stable reference between exepriments. Thus, we do not include in this analysis the ACIC 2018 dataset since R-risk* is not available due to the lack of the true propensity score.

Figure 18 Low population overlap hinders model selection for all metrics.

Figure 19 - Stacked models for the nuisances is more efficient. For each metrics the benefit of using a stacked model of linear and boosting estimators for nuisances compared to a linear model. The evaluation measure is Kendall's tau relative to the oracle R-risk* to have a stable reference between exepriments. Thus, we do not include in this analysis the ACIC 2018 dataset since R-risk* is not available due to the lack of the true propensity score.

Figure 20 - Flexible models are performant in recovering nuisances even in linear setups.

8 We obtained the dataset from https://github.com/AMLab-Amsterdam/CEVAE/tree/master/datasets/TWINS

		Strong Overlap		Weak Overlap	
Metric	Dataset	Median	IQR	Median	IQR
$\widehat{\mu-\text{risk}}$	Twins (N=11 984)	-0.32	0.12	-0.19	0.12
	ACIC 2016 (N=4 802)	-0.03	0.13	0.11	0.19
	Caussim (N=5 000)	-0.40	0.55	-0.16	0.31
	ACIC 2018 (N=5 000)	0.00	0.30	0.01	0.40
$\widehat{\mu-\text{risk}}_{IPW}$	Twins (N= 11 984)	-0.31	0.13	-0.17	0.12
	ACIC 2016 (N=4 802)	-0.02	0.13	0.11	0.19
	Caussim (N=5 000)	-0.34	0.50	0.09	0.31
	ACIC 2018 (N=5 000)	0.00	0.30	-0.01	0.43
$\widehat{\mu-\text{risk}}^{*}_{IPW}$	Twins (N= 11 984)	-0.32	0.13	-0.17	0.13
	ACIC 2016 (N=4 802)	-0.02	0.13	0.11	0.21
	Caussim (N=5 000)	-0.33	0.54	0.26	0.27
$\widehat{\tau-\text{risk}}_{IPW}$	Twins (N= 11 984)	0.13	0.12	0.27	0.12
	ACIC 2016 (N=4 802)	-0.07	0.18	0.05	0.31
	Caussim (N=5 000)	-0.19	0.43	-0.14	0.18
	ACIC 2018 (N=5 000)	-0.16	0.40	-0.11	0.66
$\widehat{\tau-\text{risk}}^{*}_{IPW}$	Twins (N= 11 984)	0.12	0.14	0.20	0.16
	ACIC 2016 (N=4 802)	-0.03	0.16	-0.09	0.43
	Caussim (N=5 000)	-0.15	0.46	-0.17	0.19
$\widehat{\text{U}-\text{risk}}$	Twins (N= 11 984)	0.13	0.12	0.02	0.25
	ACIC 2016 (N=4 802)	0.04	0.11	0.11	0.26
	Caussim (N=5 000)	0.04	0.43	-0.04	0.17
	ACIC 2018 (N=5 000)	0.12	0.26	-0.02	0.50
$\widehat{\text{U}-\text{risk}}^{*}$	Twins (N= 11 984)	0.25	0.08	-0.41	0.45
	ACIC 2016 (N=4 802)	0.08	0.13	-0.59	0.57
	Caussim (N=5 000)	0.46	0.12	0.02	0.44
$\widehat{\text{R}-\text{risk}}$	Twins (N= 11 984)	0.15	0.10	0.25	0.18
	ACIC 2016 (N=4 802)	0.07	0.12	0.22	0.15
	Caussim (N=5 000)	0.34	0.26	0.13	0.21
	ACIC 2018 (N=5 000)	0.13	0.27	0.21	0.47
$\widehat{\text{R}-\text{risk}}^{*}$	Twins (N= 11 984)	0.25	0.10	0.32	0.15
	ACIC 2016 (N=4 802)	0.12	0.12	0.25	0.15
	Caussim (N=5 000)	0.47	0.11	0.16	0.14

Table 4. Values of relative $\kappa(\ell, \tau\text{−risk})$ compared to the mean over all metrics Kendall's as shown in the boxplots of Figure 3

Selecting different seeds and parameters is crucial to draw conclusions. One strength of our study is the various number of different simulated and semi-simulated datasets. We are convinced that the usual practice of using only a small number of generation processes does not allow to draw statistically significant conclusions.

Figure 21 illustrate the dependence of the results on the generation process for caussim simulations. We highlighted the different trajectories induced by three different seeds for data generation and three different treatment ratio instead of 1000 different seeds. The result curves are relatively stable from one setup to another for $R-risk$, but vary strongly for μ-risk and μ-risk$_{IPW}$.

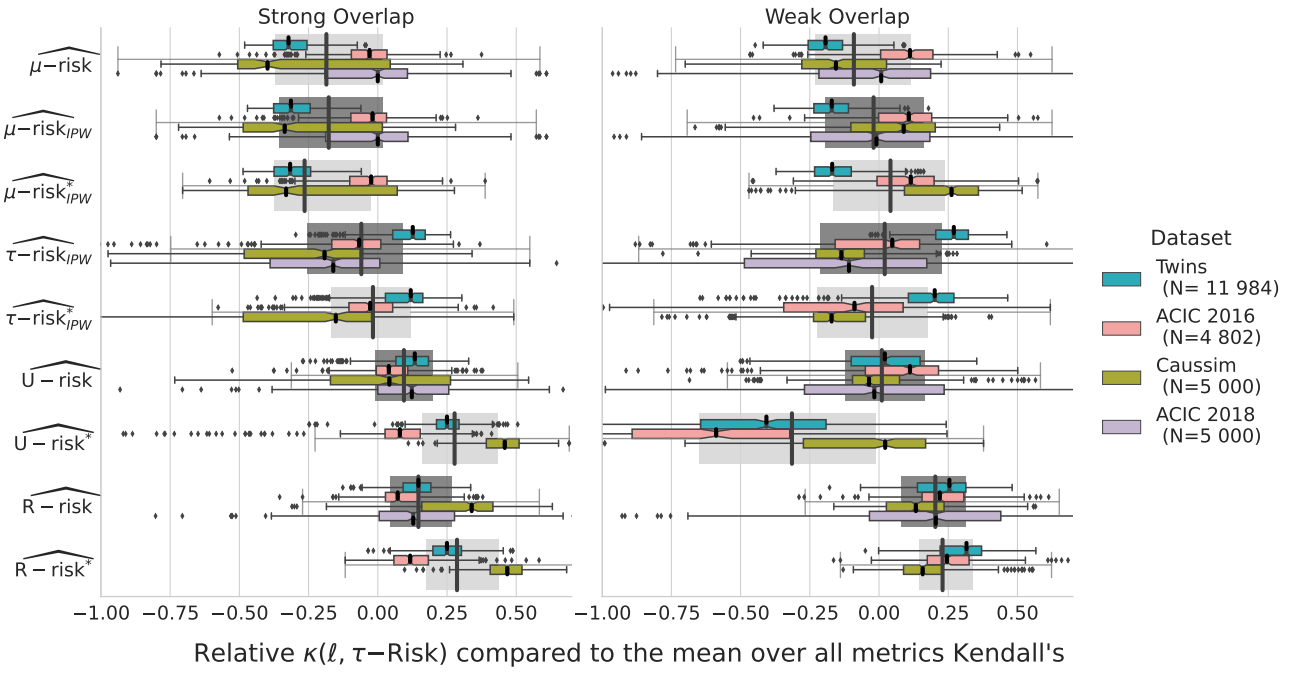


Figure 14. The R-risk is the best metric: Relative Kendall's τ agreement with τ-risk. Strong and Weak overlap correspond to the first and last tertiles of the overlap distribution measured with Normalized Total Variation eq. 17.

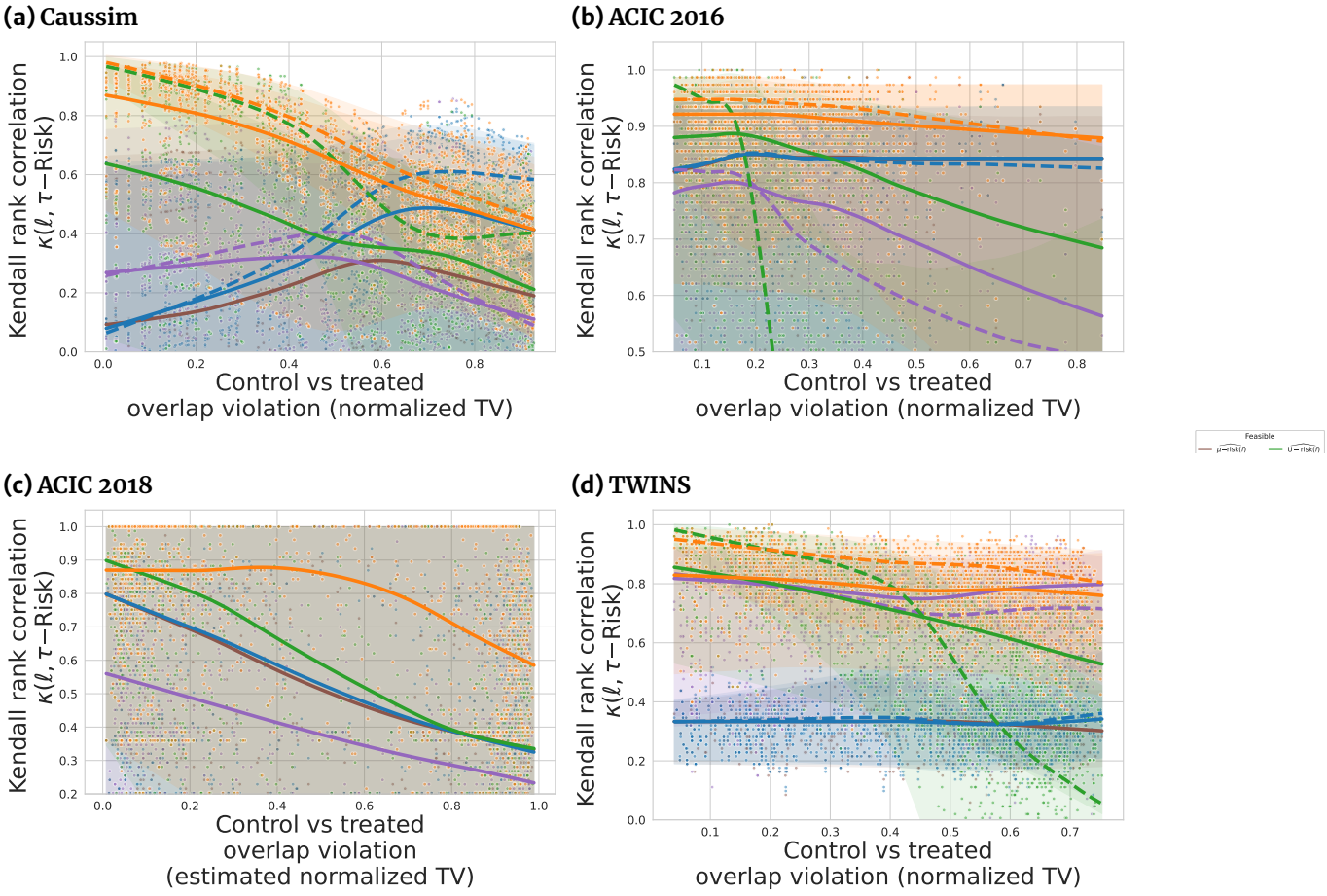


Figure 15. Agreement with τ-risk ranking of methods function of overlap violation. The lines represent medians, estimated with a lowess. The transparent bands denote the 5% and 95% confidence intervals.

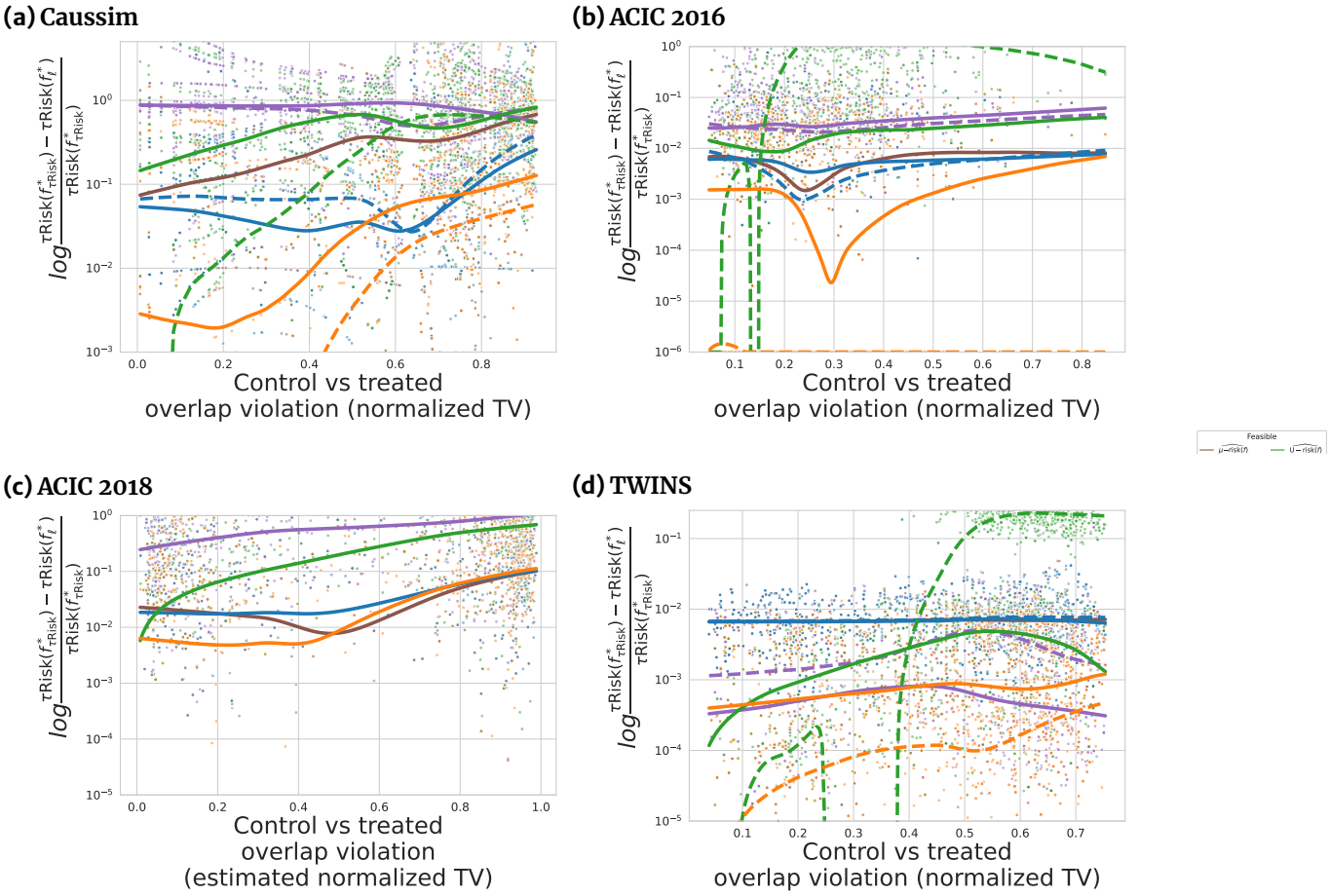


Figure 16. Metric performances by normalized tau-risk distance to the best method selected with τ-risk. All nuisances are learned with the same estimator stacking gradient boosting and ridge regression. Doted and plain lines corresponds to 60% lowess quantile estimates. This choice of quantile allows to see better the oracle metrics lines for which outliers with a value of 0 distord the curves.

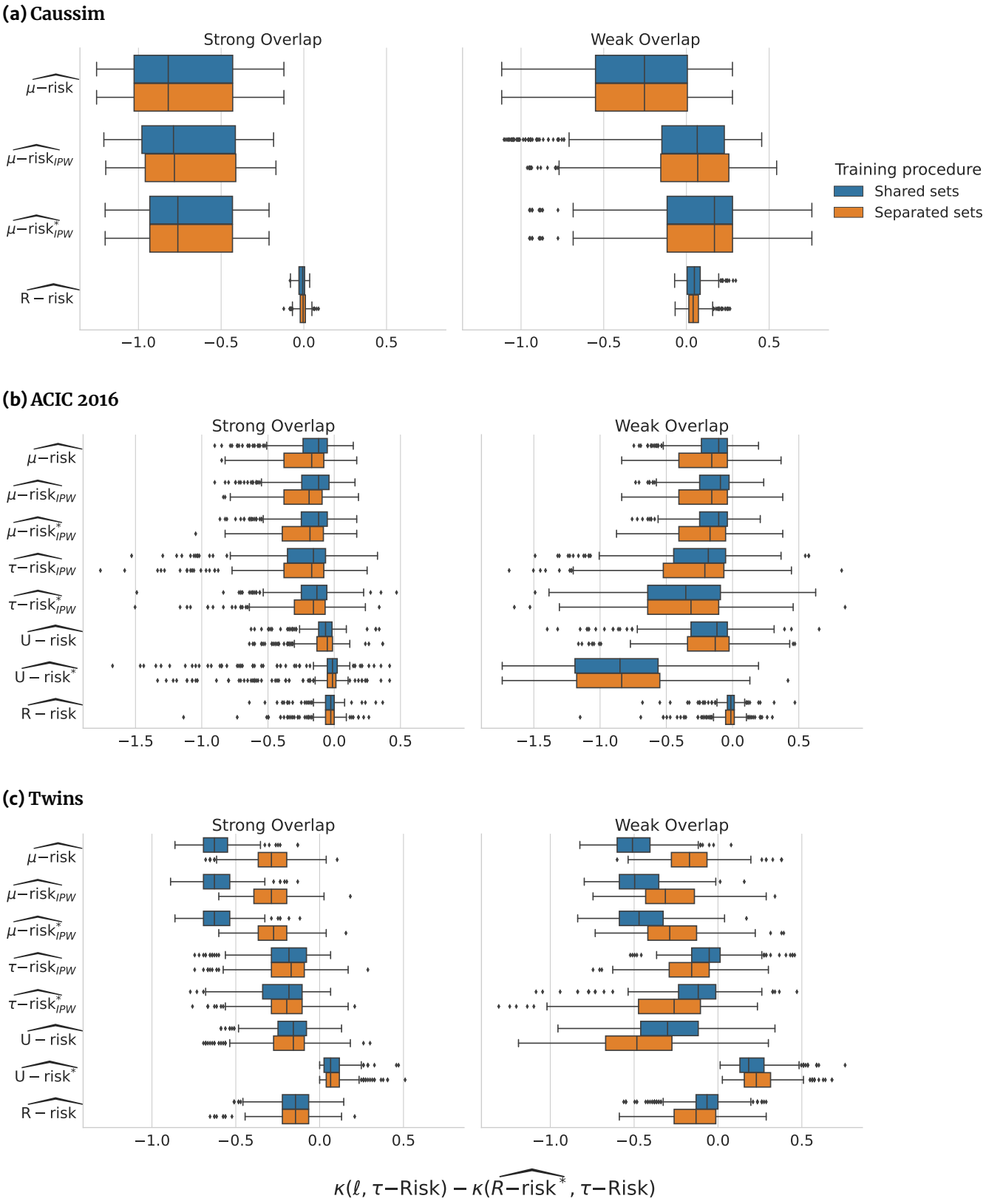


Figure 17. Results are similar between the Shared nuisances/candidate set and the Separated nuisances set procedure. The experience has not been run on the full metrics for Caussim due to computation costs.

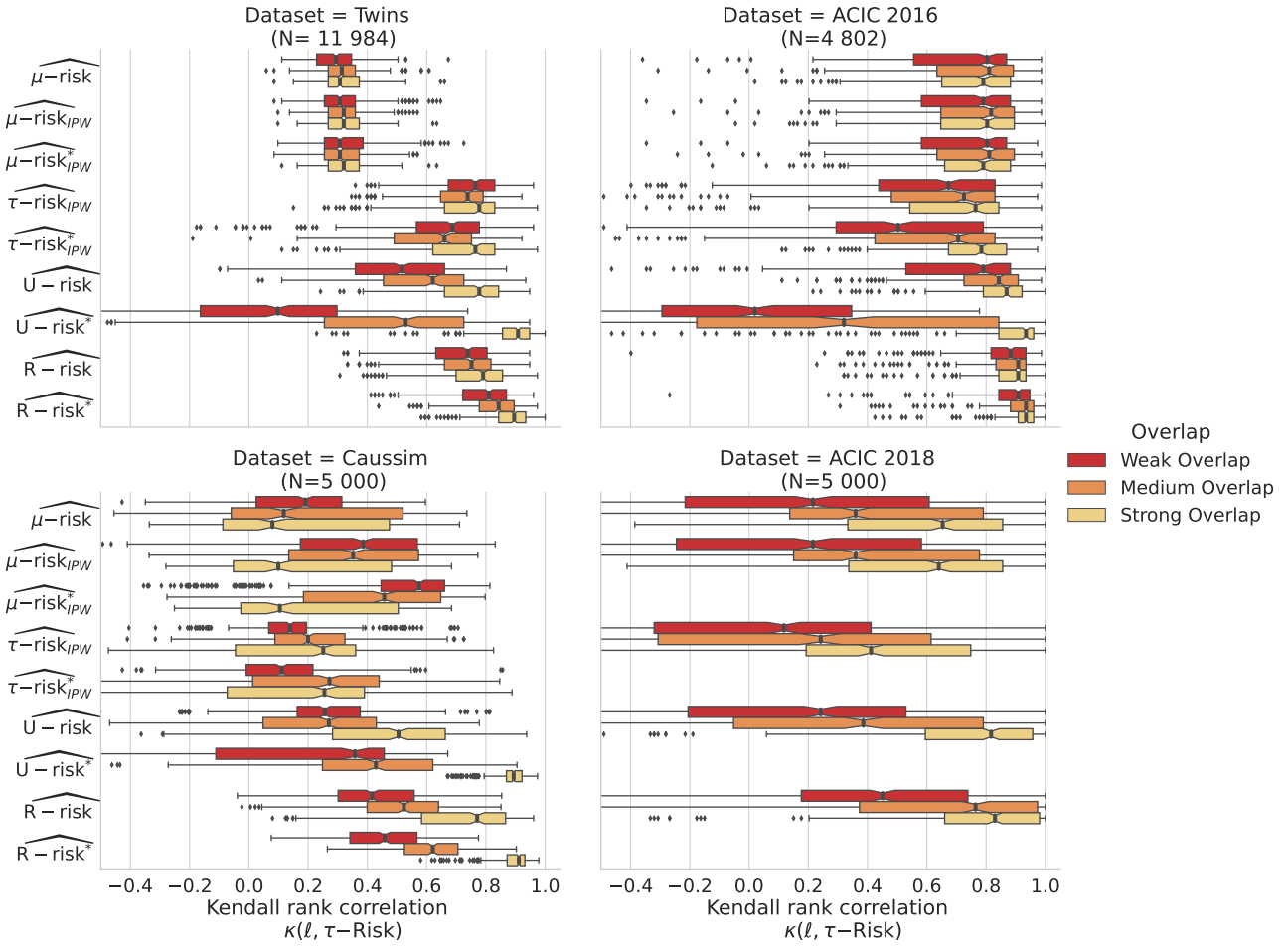


Figure 18. Low population overlap hinders causal model selection for all metrics: Kendall's τ agreement with τ-risk. Strong, medium and Weak overlap correspond to the tertiles of the overlap distribution measured with Normalized Total Varation eq. 17.

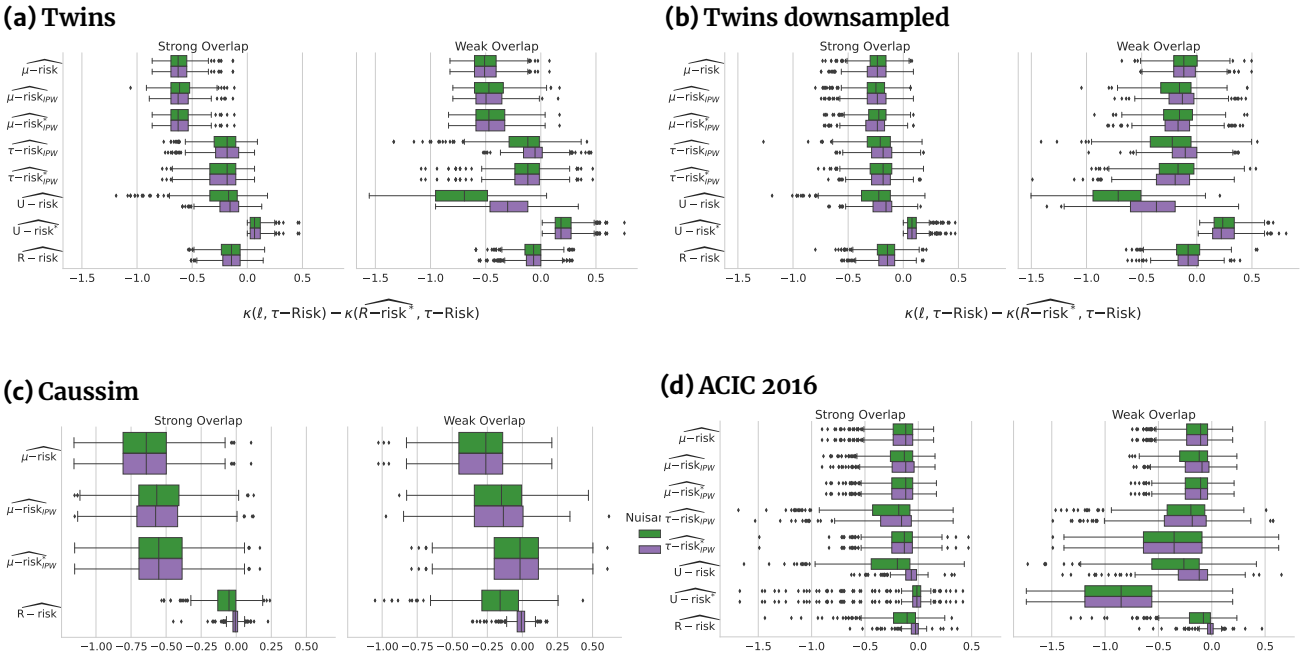


Figure 19. Learning the nuisances with stacked models (linear and gradient boosting) is important for successful model selection with R-risk. For Twins dataset, there is no improvement for stacked models compared to linear models because of the linearity of the propensity model.

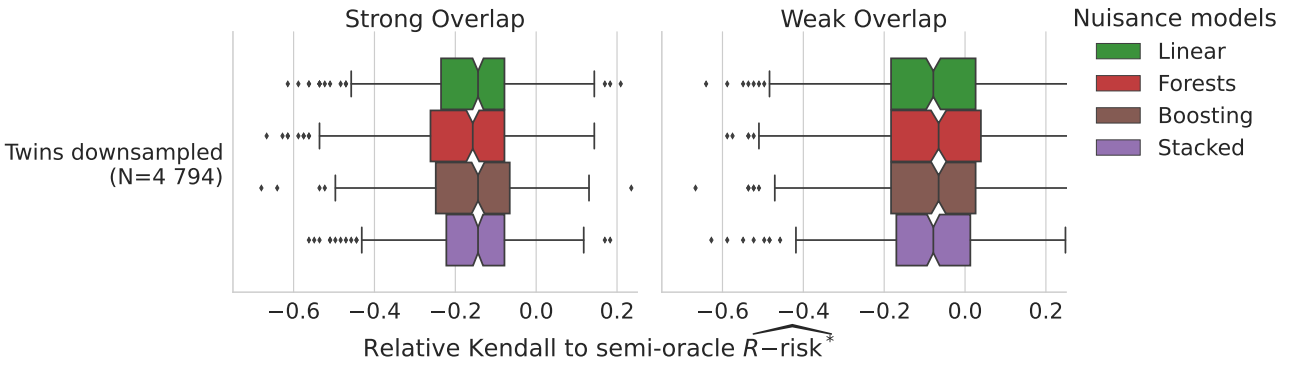
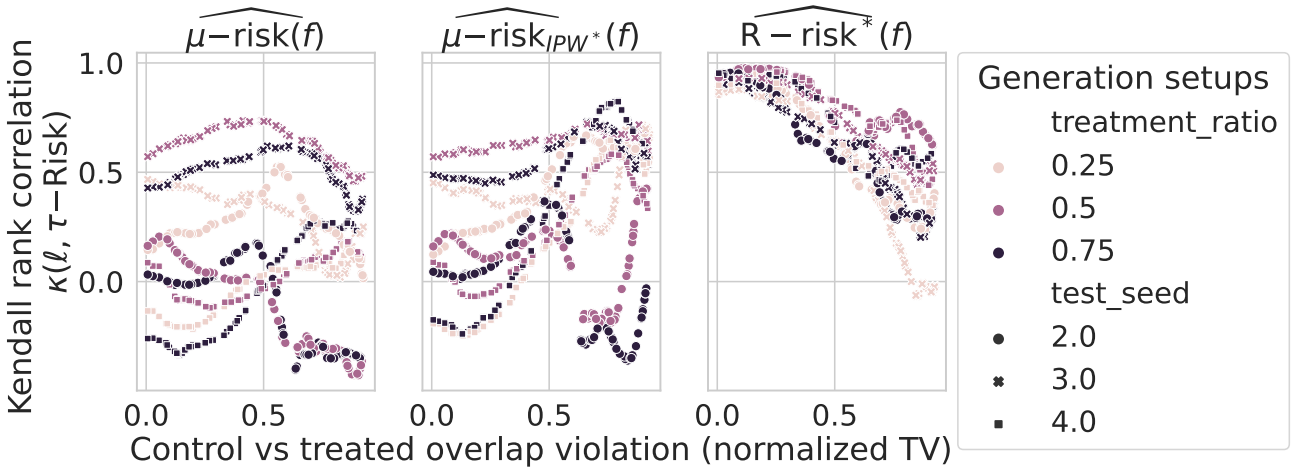


Figure 20. Flexible models are performant in recovering nuisances in the downsampled Twins dataset. The propensity score is linear in this setup, making it particularly challenging for flexible models compared to linear methods.

Figure 21. Kendall correlation coefficients for each causal metric. Each (color, shape) pair indicates a different (treatment ratio, seed) of the generation process.



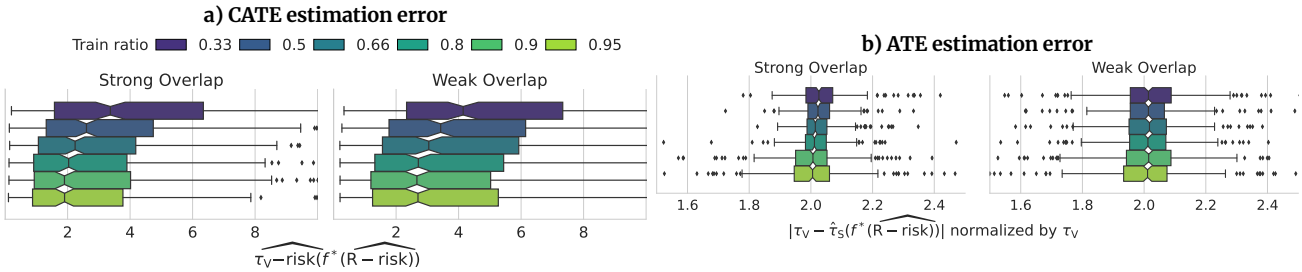


Figure 22. a) For CATE, a train/test ratio of 0.9/0.1 appears a good trade-off. b) For ATE, there is a small signal pointing also to 0.9/0.1 (K=10). for ATE. Experiences on 10 replications of all 78 instances of the ACIC 2016 data.

A.8 Data split choices

Use 90% of the data to estimate outcome models, 10% to select them

The analyst faces a compromise: given a finite data sample, should she allocate more data to estimate the outcome model, thus improving the quality of the outcome model but leaving little data for model selection. Or, she could choose a bigger test set for model selection and effect estimation. For causal model selection, there is no established practice (as reviewed in A.8).

We investigate such tradeoff varying the ratio between train and test data size. For this, we first split out 30% of the data as a holdout set $\mathcal{V}$ on which we use the oracle response functions to derive silver-standard estimates of causal quantities. We then use the standard estimation procedure on the remaining 70% of the data, splitting it into train $\mathcal{T}$ and test $\mathcal{S}$ of varying sizes. We finally measure the error between this estimate and the silver standard.

We consider two different analytic goals: estimating a average treatment effect –a single number used for policy making– and a CATE –a full model of the treatment effect as a function of covariates X. Given that the latter is a much more complex object than the former, the optimal train/test ratio might vary. To measure errors, we use for the ATE the relative absolute ATE bias between the ATE computed with the selected outcome model on the test set, and the true ATE as evaluated on the holdout set $\mathcal{V}$. For the CATE, we compare the τ-risk of the best selected model applied on the holdout set $\mathcal{V}$. We explore this trade-off for the ACIC 2016 dataset and the R-risk.

Figure 22 shows that a train/test ratio of 0.9/0.1 (K=10) or 0.8/0.2 (K=5) appears best to estimate CATE and ATE.

Heterogeneity in practices for data split

Splitting the data is common when using machine learning for causal inference, but practices vary widely in terms of the fraction of data to allocate to train models, outcomes and nuisances, and to evaluate them.

Before even model selection, data splitting is often required for estimation of the treatment effect, ATE or CATE, for instance to compute the nuisances required to optimize the outcome model (as the R-risk, Definition 5). The most frequent choice is use 80% of the data to fit the models, and 20% to evaluate them. For instance, for CATE estimation, the R-learner has been introduced using K-folds with K = 5 and K = 10: 80% of the data (4 folds) to train the nuisances and the remaining fold to minimize the corresponding R-loss [42]. Yet, it has been implemented with K=5 in causallib [93] or K=3 in econML [94]. Likewise, for ATE estimation, [43] introduce doubly-robust machine learning, recommending K=5 based on an empirical comparison K=2. However, subsequent works use doubly robust ML with varying choices of K: [95] use K=3, [96] use K=2. In the econML implementation, K is set to 3 [94]. [97] evaluate various machine-learning approaches –including R-learners– using K=5 and 10, drawing inspiration from the TMLE literature which sets K=5 in the TMLE package [98].

Causal model selection has been much less discussed. The only study that we are aware of, [50], use a different data split: a 2-folds train/test procedure, training the nuisances on the first half of the data, and using the second half to estimate the R-risk and select the best treatment effect model.

A.9 Sensitivity of the results to different ratio of causal effect to baseline response

Going beyond overlap, we study the effect of another important parameter of the data generation process: the ratio of causal effect to the baseline response.

For each dataset of our simulation (caussim), we measure this ratio as the absolute mean difference between the causal effect and the baseline response, measured empirically on each dataset instance $\Delta_\mu = \frac{1}{N}\sum_{i=1}^{N} \left| \frac{\mu_1(x_i) - \mu_0(x_i)}{\mu_0(x_i)} \right|$. We vary a dedicated parameter (ω in Algorithm 2) in the simulation to force this ratio to cover a wide range of values. For 179 dataset instances, it ranges several order of magnitudes with the full distribution of values given in Table 5.

	count	mean	std	min	1%	10%	25%	50%	75%	90%	99%	max
effect_ratio	179	159.6	852.4	0.004674	0.004765	0.02483	0.1566	1.879	21.12	135.1	3394	9675

Table 5. Distribution of the causal effect ratio Δ_μ for the experiment on causal effect ratio, results in Figure 7.

References

1. Karel GM Moons, Patrick Royston, Yvonne Vergouwe, Diederick E Grobbee, and Douglas G Altman. Prognosis and prognostic research: what, why, and how? *Bmj*, 338, 2009.
2. Ewout W Steyerberg. *Clinical prediction models*. Springer.
3. Andrew L Beam and Isaac S Kohane. Big data and machine learning in health care. *Jama*, 319(13):1317–1318, 2018.
4. Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.
5. M Khojaste-Sarakhsi, Seyedhamidreza Shahabi Haghighi, SMT Fatemi Ghomi, and Elena Marchiori. Deep learning for alzheimer's disease diagnosis: A survey. *Artificial Intelligence in Medicine*, page 102332, 2022.
6. Zhenwei Zhang and Ervin Sejdić. Radiological images and machine learning: trends, perspectives, and prospects. *Computers in biology and medicine*, 108:354–370, 2019.
7. Adam Yala, Constance Lehman, Tal Schuster, Tally Portnoi, and Regina Barzilay. A deep learning mammography-based model for improved breast cancer risk prediction. *Radiology*, 292(1):60–66, 2019.
8. Li Shen, Laurie R Margolies, Joseph H Rothstein, Eugene Fluder, Russell McBride, and Weiva Sieh. Deep learning to improve breast cancer detection on screening mammography. *Scientific reports*, 9(1):12495, 2019.
9. Ali Bou Nassif, Manar Abu Talib, Qassim Nasir, Yaman Afadar, and Omar Elgendy. Breast cancer detection using artificial intelligence techniques: A systematic literature review. *Artificial Intelligence in Medicine*, page 102276, 2022.
10. Stephen J Mooney and Vikas Pejaver. Big data in public health: terminology, machine learning, and privacy. *Annual review of public health*, 39:95, 2018.
11. Rishi J Desai, Shirley V Wang, Muthiah Vaduganathan, Thomas Evers, and Sebastian Schneeweiss. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA network open*, 3(1):e1918962–e1918962, 2020.
12. Gregory E Simon, Eric Johnson, Jean M Lawrence, Rebecca C Rossom, Brian Ahmedani, Frances L Lynch, Arne Beck, Beth Waitzfelder, Rebecca Ziebell, Robert B Penfold, et al. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. *American Journal of Psychiatry*, 175(10):951–960, 2018.
13. Steven Horng, David A Sontag, Yoni Halpern, Yacine Jernite, Nathan I Shapiro, and Larry A Nathanson. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. *PloS one*, 12(4):e0174708, 2017.
14. Hanyin Wang, Yikuan Li, Seema A Khan, and Yuan Luo. Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. *Artificial intelligence in medicine*, 110:101977, 2020.
15. Irena Spasic, Goran Nenadic, et al. Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3):e17984, 2020.
16. Douglas G Altman, Yvonne Vergouwe, Patrick Royston, and Karel GM Moons. Prognosis and prognostic research: validating a prognostic model. *Bmj*, 338, 2009.
17. Russell A Poldrack, Grace Huckins, and Gael Varoquaux. Establishment of best practices for evidence for prediction: a review. *JAMA psychiatry*, 77(5):534–540, 2020.
18. Gaël Varoquaux and Olivier Colliot. Evaluating machine learning models and their diagnostic value, 2022.
19. Jeremy C Wyatt and Douglas G Altman. Commentary: Prognostic models: clinically useful or quickly forgotten? *Bmj*, 311(7019): 1539–1541, 1995.
20. Mark Alan Fontana, Stephen Lyman, Gourab K Sarker, Douglas E Padgett, and Catherine H MacLean. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? *Clinical orthopaedics and related research*, 477(6):1267, 2019.
21. Jonathan M. Snowden, Sherri Rose, and Kathleen M. Mortimer. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology*, 173(7):731–738, April 2011.
22. Matthew Sperrin, David Jenkins, Glen P Martin, and Niels Peek. Explicit causal reasoning is needed to prevent prognostic models being victims of their own success. *Journal of the American Medical Informatics Association*, 26(12):1675–1676, 2019.
23. Tony Blakely, John Lynch, Koen Simons, Rebecca Bentley, and Sherri Rose. Reflection on modern methods: when worlds collide–prediction, machine learning and causal inference. *International journal of epidemiology*, 49(6):2058–2064, 2020.
24. MA Hernán and JM Robins. *Causal Inference: What If*. CRC Boca Raton, FL.
25. Tyler J VanderWeele. Principles of confounder selection. *European journal of epidemiology*, 34:211–219, 2019.
26. James M. Robins and Sander Greenland. The role of model selection in causal inference from non experimental data. *American Journal of Epidemiology*, 123(3):392–402.
27. T. Wendling, K. Jung, A. Callahan, A. Schuler, N. H. Shah, and B. Gallego. Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. *Statistics in Medicine*, (23):3309–3324.
28. Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55.
29. Peter C. Austin and Elizabeth A. Stuart. 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*, 34(28):3661–3679.
30. Sabrina Casucci, Li Lin, Sharon Hewner, and Alexander Nikolaev. Estimating the causal effects of chronic disease combinations on 30-day hospital readmissions based on observational medicaid data. *Journal of the American Medical Informatics Association*, 25(6): 670–678, 2018.
31. Elysia Grose, Samuel Wilson, Jeffrey Barkun, Kimberly Bertens, Guillaume Martel, Fady Balaa, and Jad Abou Khalil. Use of propensity score methodology in contemporary high-impact surgical literature. *Journal of the American College of Surgeons*, 230(1):101–112.e2.
32. Xiaogang Su, Annette T Peña, Lei Liu, and Richard A Levine. Random forests of interaction trees for estimating individualized treatment effects in randomized trials. *Statistics in medicine*, 37(17):2547–2560, 2018.
33. Andrea Lamont, Michael D Lyons, Thomas Jaki, Elizabeth Stuart, Daniel J Feaster, Kukatharmini Tharmaratnam, Daniel Oberski, Hemant Ishwaran, Dawn K Wilson, and M Lee Van Horn. Identification of predicted individual treatment effects in randomized clinical trials. *Statistical methods in medical research*, 27(1):142–157, 2018.

34. Jeroen Hoogland, Joanna IntHout, Michail Belias, Maroeska M Rovers, Richard D Riley, Frank E. Harrell Jr, Karel GM Moons, Thomas PA Debray, and Johannes B Reitsma. A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. *Statistics in medicine*, 40(26):5961–5981, 2021.
35. Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, (1):217–240.
36. Mark J. van der Laan and Sherri Rose. *Targeted Learning*. Springer Series in Statistics.
37. Megan S. Schuler and Sherri Rose. Targeted maximum likelihood estimation for causal inference in observational studies. *American Journal of Epidemiology*, 185(1):65–73.
38. Scott Powers, Junyang Qian, Kenneth Jung, Alejandro Schuler, Nigam H. Shah, Trevor Hastie, and Robert Tibshirani. Some methods for heterogeneous treatment effect estimation in high dimensions. *Statistics in Medicine*, 37(11):1767–1787.
39. Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.
40. Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178.
41. Sören R. Künzel, Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the National Academy of Sciences*, 116(10):4156–4165.
42. Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319.
43. Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, page 71.
44. Gang Fang, Izabela E Annis, Jennifer Elston-Lafata, and Samuel Cykert. Applying machine learning to predict real-world individual treatment effects: insights from a virtual patient cohort. *Journal of the American Medical Informatics Association*, 26(10):977–988, 2019.
45. Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.
46. Florent Le Borgne, Arthur Chatton, Maxime Léger, Rémi Lenain, and Yohann Foucher. G-computation and machine learning for estimating the causal effects of binary exposure statuses on binary outcomes. *Scientific reports*, 11(1):1435, 2021.
47. Jinma Ren, Paul Cislo, Joseph C Cappelleri, Patrick Hlavacek, and Marco DiBonaventura. Comparing g-computation, propensity score-based weighting, and targeted maximum likelihood estimation for analyzing externally controlled trials with both measured and unmeasured confounders: a simulation study. *BMC Medical Research Methodology*, 23(1):18, 2023.
48. Jean Kaddour, Aengus Lynch, Qi Liu, Matt J Kusner, and Ricardo Silva. Causal machine learning: A survey and open problems. arxiv 2022. *arXiv preprint arXiv:2206.15475*, 2022.
49. Victor Chernozhukov, Christian Hansen, Nathan Kallus, Martin Spindler, and Vasilis Syrgkanis. Applied causal inference powered by ml and ai. *arXiv preprint arXiv:2403.02467*, 2024.
50. Alejandro Schuler, Michael Baiocchi, Robert Tibshirani, and Nigam Shah. A comparison of methods for model selection when estimating individual treatment effects. *arXiv:1804.05146 [cs, stat]*.
51. Ahmed Alaa and Mihaela Van Der Schaar. Validating causal inference models via influence functions. *International Conference on Machine Learning*, pages 191–201.
52. Mary E. Charlson, Peter Pompei, Kathy L. Ales, and C.Ronald MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. *Journal of Chronic Diseases*, 40(5):373–383.
53. James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7(9):1393–1512.
54. Ashley I Naimi and Brian W Whitcomb. Defining and identifying average treatment effects. *American Journal of Epidemiology*, 2023.
55. Guido W. Imbens and Donald B. Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.
56. Donald B Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100(469):322–331.
57. P. M. Robinson. Root-n-consistent semiparametric regression. *Econometrica*, (4):931–954.
58. Mark J van der Laan, MJ Laan, and JM Robins. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media.
59. Peter Schulam and Suchi Saria. Reliable decision support using counterfactual models. *Advances in neural information processing systems*, 30.
60. Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.
61. Chanelle J. Howe, Stephen R. Cole, Daniel J. Westreich, Sander Greenland, Sonia Napravnik, and Joseph J. Eron. Splines for trend analysis and continuous confounder control. 22(6):874–875.
62. Aris Perperoglou, Willi Sauerbrei, Michal Abrahamowicz, and Matthias Schmid. A review of spline function procedures in r. 19(1):46.
63. Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in Neural Information Processing Systems*, volume 20.
64. Lingjie Shen, Gijs Geleijnse, and Maurits Kaptein. Rctrep: An r package for the validation of estimates of average treatment effects. *Journal of Statistical Software*, 2023.
65. Kenneth R. Niswander and United States National Institute of Neurologiacal Diseases and Stroke. *The Women and Their Pregnancies: The Collaborative Perinatal Study of the National Institute of Neurological Diseases and Stroke*. National Institute of Health. Google-Books-ID: AobdVhlhDQkC.
66. Yishai Shimoni, Chen Yanover, Ehud Karavani, and Yaara Goldschmnidt. Benchmarking framework for performance-evaluation of causal inference analysis. *arXiv:1802.05046 [cs, stat]*.
67. M. F. MacDorman and J. O. Atkinson. Infant mortality statistics from the linked birth/infant death data set–1995 period data. *Monthly Vital Statistics Report*, 46(6 Suppl 2):1–22, February 1998.
68. Christos Louizos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. *Advances in neural information processing systems*.
69. Douglas Almond, Kenneth Y. Chay, and David S. Lee. The Costs of Low Birth Weight. *The Quarterly Journal of Economics*, 120(3), 2005.
70. M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, June 1938.
71. Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

72. Pierre Gutierrez and Jean-Yves Gerardy. Causal inference and uplift modeling a review of the literature. *Proceedings of The 3rd International Conference on Predictive Applications and APIs*, (67):14.
73. John C. Platt and John C. Platt. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers*, pages 61–74, 1999.
74. Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. page 8.
75. Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pages 625–632. ACM Press, 2005.
76. Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the Calibration of Modern Neural Networks. *Advances in Neural Information Processing Systems*, 34:15682–15694, 2021.
77. Alexandre Perez-Lebel, Marine Le Morvan, and Gaël Varoquaux. Beyond calibration: estimating the grouping loss of modern neural networks. *arXiv preprint arXiv:2210.16315*, 2022.
78. Xavier Bouthillier, Pierre Delaunay, Mirko Bronzi, Assya Trofimov, Brennan Nichyporuk, Justin Szeto, Nazanin Mohammadi Sepahvand, Edward Raff, Kanika Madan, Vikram Voleti, Samira Ebrahimi Kahou, Vincent Michalski, Tal Arbel, Chris Pal, Gael Varoquaux, and Pascal Vincent. Accounting for variance in machine learning benchmarks. *Proceedings of Machine Learning and Systems*, 3:747–769.
79. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830.
80. Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.
81. Craig A. Rolling and Yuhong Yang. Model selection for estimating treatment effects. 76(4):749–769.
82. Yuta Saito and Shota Yasui. Counterfactual cross-validation: Stable model selection procedure for causal inference models. In *International Conference on Machine Learning*, pages 8398–8407. PMLR, 2020.
83. Fredrik D Johansson, Uri Shalit, Nathan Kallus, and David Sontag. Generalization bounds and representation learning for estimation of potential outcomes and causal effects. *The Journal of Machine Learning Research*, 23(1):7489–7538, 2022.
84. Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.
85. Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. *Advances in Neural Information Processing Systems*, 33:11637–11649.
86. Peter C. Austin. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research*, (3):399–424, May 2011.
87. Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773.
88. Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On integral probability metrics, \phi-divergences and binary classification. *arXiv:0901.2698 [cs, math]*.
89. Alicia Curth, David Svensson, and James Weatherall. Really doing great at estimating CATE? a critical look at ML benchmarking practices in treatment effect estimation. *Neurips Process 2021*, page 14.
90. Mark J. van der Laan, Eric C. Polley, and Alan E. Hubbard. Super learner. 6.
91. Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823. PMLR, 2015.
92. Edward L Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311.
93. Yishai Shimoni, Ehud Karavani, Sivan Ravid, Peter Bak, Tan Hung Ng, Sharon Hensley Alford, Denise Meade, and Yaara Goldschmidt. An evaluation toolkit to guide model selection and cohort definition in causal inference. *arXiv preprint arXiv:1906.00442*, 2019.
94. Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML.
95. Nicolas Loiseau, Paul Trichelair, Maxime He, Mathieu Andreux, Mikhail Zaslavskiy, Gilles Wainrib, and Michael G. B. Blum. External control arm analysis: an evaluation of propensity score approaches, G-computation, and doubly debiased machine learning. *BMC Medical Research Methodology*, 22, 2022.
96. Zijun Gao, Trevor Hastie, and Robert Tibshirani. Assessment of heterogeneous treatment effect estimation accuracy via matching. *Statistics in Medicine*, (17), 2021.
97. Ashley I Naimi, Alan E Mishler, and Edward H Kennedy. Challenges in obtaining valid causal effect estimates with machine learning algorithms. *American Journal of Epidemiology*, 2021.
98. Susan Gruber and Mark J. van der Laan. tmle: An R package for targeted maximum likelihood estimation. *Journal of Statistical Software*, 51(13):1–35, 2012.