Supplementary materials:

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

A Variability of ATE estimation on ACIC 2016

Figure S1 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[11]. Figure S1 shows large variations obtained across different outcome estimators on semi-synthetic datasets[18]. 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[54] 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 S1: Hyper-parameters grid used for ACIC 2016 ATE variability

B 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[59], where the oracle CATE is known, Schuler et al. [70] compare four causal risks, concluding that for CATE estimation the best model-selection risk is the so-called R-risk[52] –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, 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[1] studied empirically both MSE and reweighted MSE risks on the semi-synthetic ACIC 2016 datasets[18], 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 influence of overlap, an important parameter of the data generation process which makes a given causal metric appropriate[16]. 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[65], an IPW estimate[25], a doubly robust estimator[68], or debiasing the error with influence functions[1]. 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

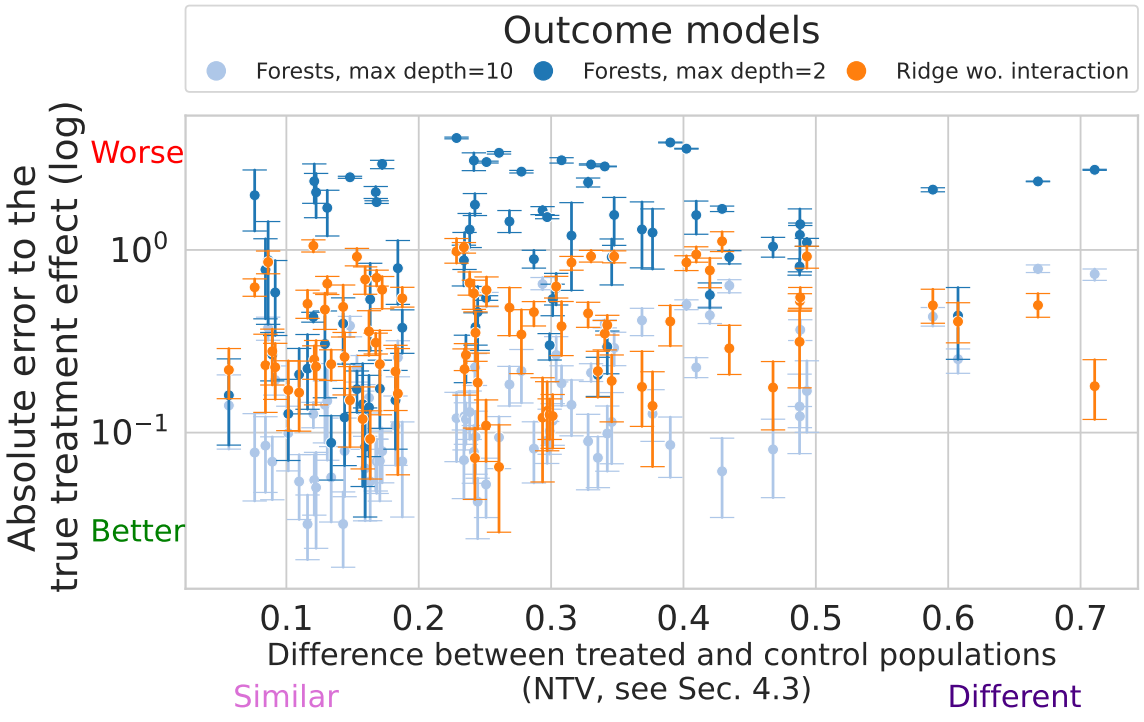


Figure S1: Different outcome models lead to different estimation errors on the Average Treatment Effects, on 77 classic simulations with known true causal effect[18]. The different models are ridge regression and random forests with different hyper-parameters (details A). The different configurations are plotted as a function of increasing difference between treated and untreated population –see section . There is no systematic best performer; data-driven model selection is important.

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)[40,71] or Double Machine Learning[14]. Here also, theories require asymptotic regimes and models to be *well-specified*.

By contrast, Johansson et al. [34] studies causal estimation without assuming that estimators are well specified. They derive 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. 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[70].

C Causal assumptions

We assume the following four assumptions, referred as strong ignorability and necessary to assure identifiability of the causal estimands with observational data[67]:

Assumption 1 (Unconfoundedness)

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

This condition –also called ignorability– is equivalent to the conditional independence on $e(X)$[66]: $\{Y(0), Y(1)\} \perp\!\!\!\perp A|e(X)$.

Assumption 2 (Overlap, also known as Positivity))

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

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[16], 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.

Assumption 3 (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[26].

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*[33].

D Definitions of feasible risks

Definition 1 (Factual μ-risk) [72] *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$^\star_{IPW}$) [38] *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,*

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

Definition 3 (τ-risk$^\star_{IPW}$) [87] *The CATE $\tau(x)$ can be estimated with a regression against inverse propensity weighted outcomes*[4,25,87]*, the τ-risk$_{IPW}$.*

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

Definition 4 (U-risk*) [37,52] *Based on the Robinson decomposition –eq. 7, the U-learner uses the $A - e(X)$ term in the denominator. The derived risk is:*

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

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*) [52,70] *The R-risk also uses two nuisance m and e:*

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

It is also based on the Robinson decomposition –eq. 7.

E Proofs: Links between feasible and oracle risks

E.1 Reformulation of the R-risk as reweighted τ-risk

Proposition 1 (R-risk as reweighted τ-risk) ***Proof 1*** *We consider the R-decomposition:*[64]*,*

$$y(a) = m(x) + \big(a - e(x)\big)\tau(x) + \varepsilon(x; a) \tag{11}$$

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

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

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

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

The second term is null since, $\mathbb{E}[\varepsilon(x,a)|X,A] = 0$.

The third term corresponds to the modulated residuals : $\tilde{\sigma}_B^2(0) + \tilde{\sigma}_B^2(1)$

E.2 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:

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

However, we still can have large differences between τ-risk and μ-risk$_{IPW}$ coming from heterogeneous errors between populations as shown experimentally in Schuler et al. [70] 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[70] 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:

$$\tau\text{-risk}(\mu) = \mathbb{E}_{X\sim p(X)}[(\tau(X) - \tau_\mu(X))^2] = 0 \tag{12}$$

$$\mu\text{-risk}(\mu) = \mathbb{E}_{(Y,X,A)\sim p(Y;X;A)}[\big(Y - \mu_A(X)\big)^2] \tag{13}$$
$$= \int_{\mathcal{X},\mathcal{A}} \varepsilon(x,a)^2 p(a \mid x)\, p(x)\, dx\, da \leq \sigma_B^2(0) + \sigma_B^2(1)$$

$$\mu\text{-risk}_{IPW}(\mu) = \sigma_B^2(0) + \sigma_B^2(1) \quad \text{from Lemma ??} \tag{14}$$

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

from Proposition 1 (15)

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.

F 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[6,7]. 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 $\mathbb{P}(X|A=1)$[16,34]. 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[22] is typically used in the context of causal inference[72,34]. 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[81] 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 $\mathbb{P}(X|A)$:

$$\widehat{NTV}(e, 1-e) = \frac{1}{2N}\sum_{i=1}^{N} \left| \frac{e(x_i)}{p_A} - \frac{1-e(x_i)}{1-p_A} \right| \tag{16}$$

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 S2.

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 S4,
- Verifying that setups with penalized overlap from ACIC 2016 have a higher total variation distance than unpenalized setups in Figure S3.

- Verifying that the Inverse Propensity Weights extrema (the inverse of the ν overlap constant appearing in the overlap Assumption 2) positevely correlates with NTV for Caussim, ACIC 2016 and Twins in Figure S5. 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 S2 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.

G Experiments

G.1 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[60]). We carefully control the amount of overlap between treated and control populations, a crucial assumption for causal inference. Figure S6 illustrates 2D examples of the simulation.

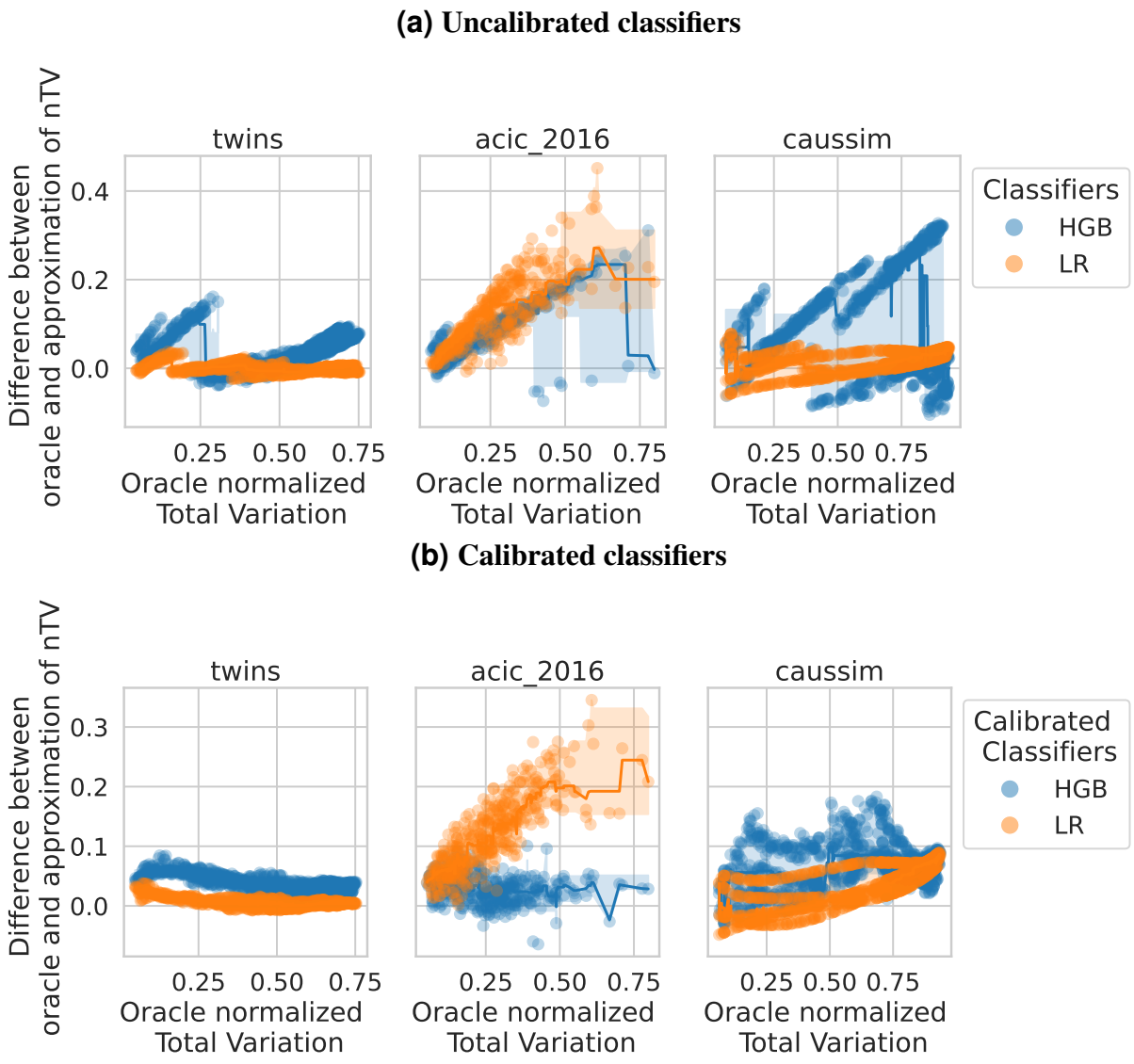


Figure S2: a) Without calibration, estimation of NTV is not trivial even for boosting models. b) Calibrated classifiers are able to recover the true Normalized Total Variation for all datasets where it is available.

Figure S3: NTV recovers well the overlap settings described in the ACIC paper[18]

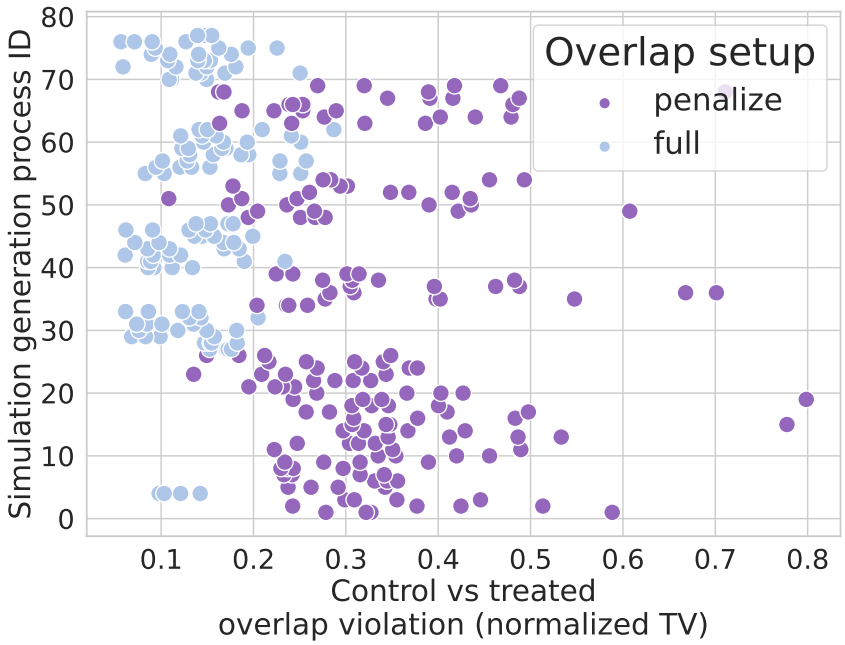
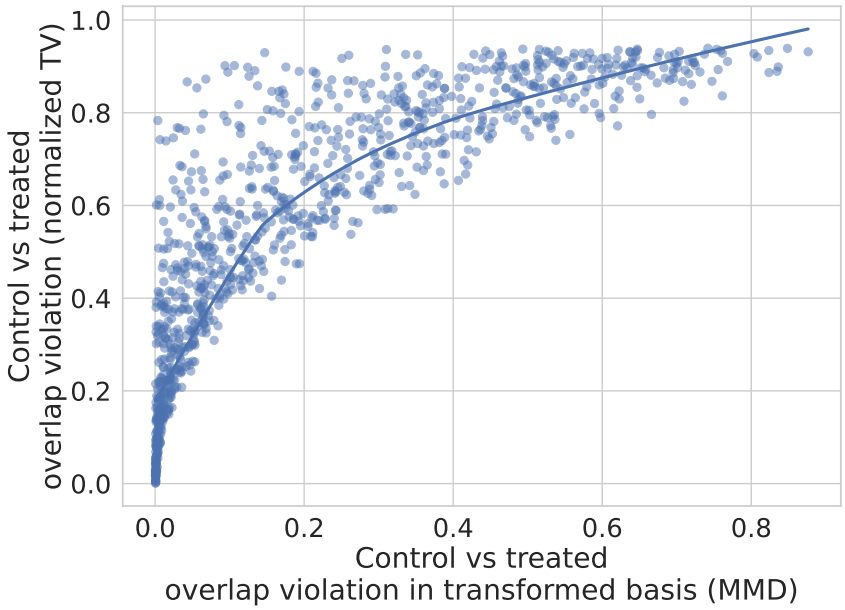


Figure S4: Good correlation between overlap measured as normalized Total Variation and Maximum Mean Discrepancy (200 sampled Caussim datasets)



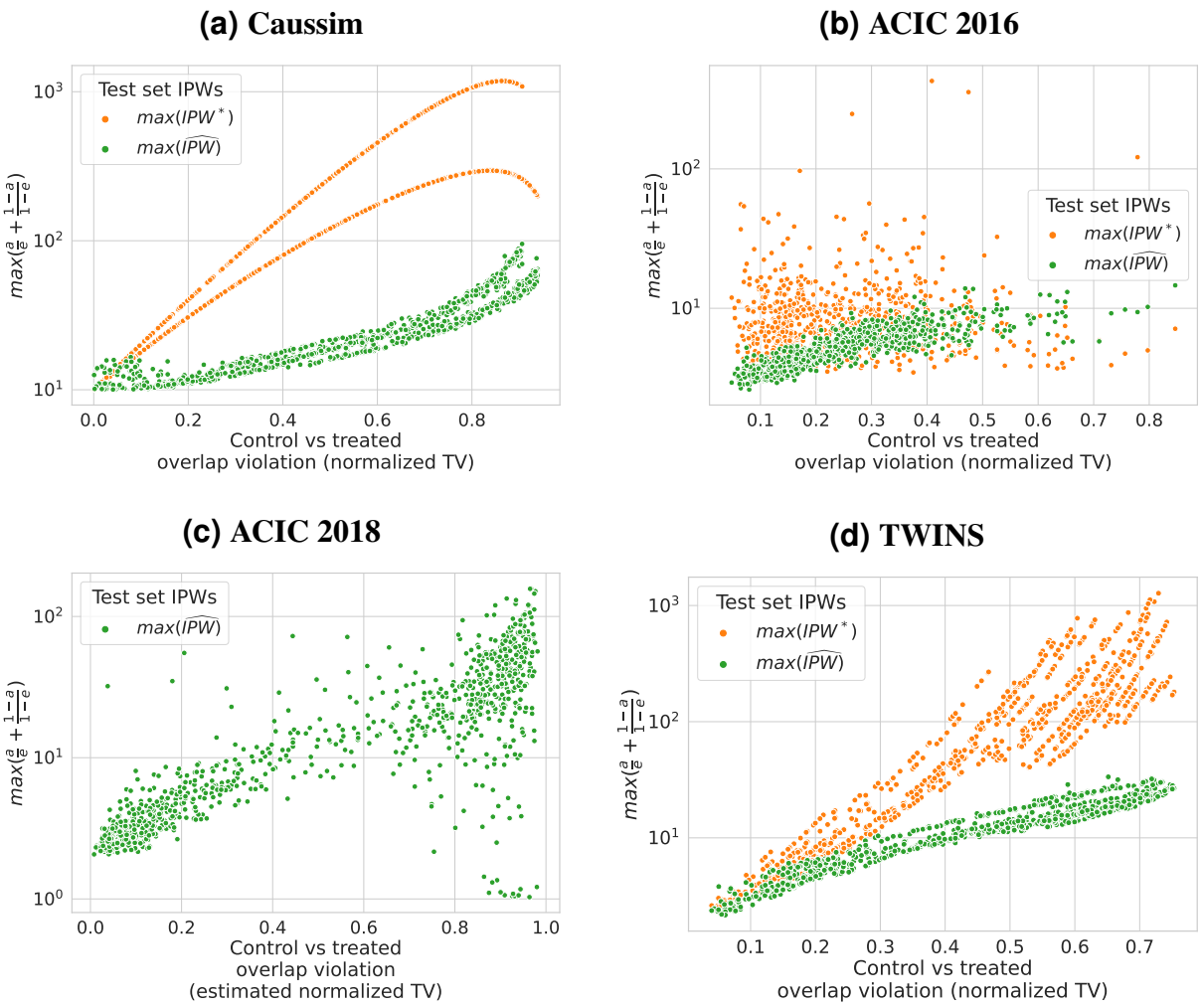


Figure S5: Maximal value of Inverse Propensity Weights increases exponentially with the overlap as measure by Normalized Total Variation.

- 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:

$$\mathbb{P}(x|A = a) = W \cdot \mathcal{N}\Big(\begin{bmatrix} (1-2a)\theta \\ 0 \end{bmatrix} ; \begin{bmatrix} \sigma_0 & 0 \\ 0 & \sigma_1 \end{bmatrix} \Big) \tag{17}$$

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 **??**). 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 $RBF(x, \cdot)\ with\ x \sim \mathbb{P}(x)$ with these representers: $z(x) = [RBF_\gamma(x, b_d)]_{d=1..D} \cdot Z^T \in \mathbb{R}^D$ 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:

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

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

- Adding a Gaussian noise, $\varepsilon \sim \mathcal{N}(0, \sigma(x; a))$, we construct the potential outcomes: $y(a) = \mu_0(x) + a\,\tau(x) + \varepsilon(x, a)$

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

G.2 Details on the semi-simulated datasets

ACIC 2018 [18]: 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

[5] We use the Sklearn implementation,[54]

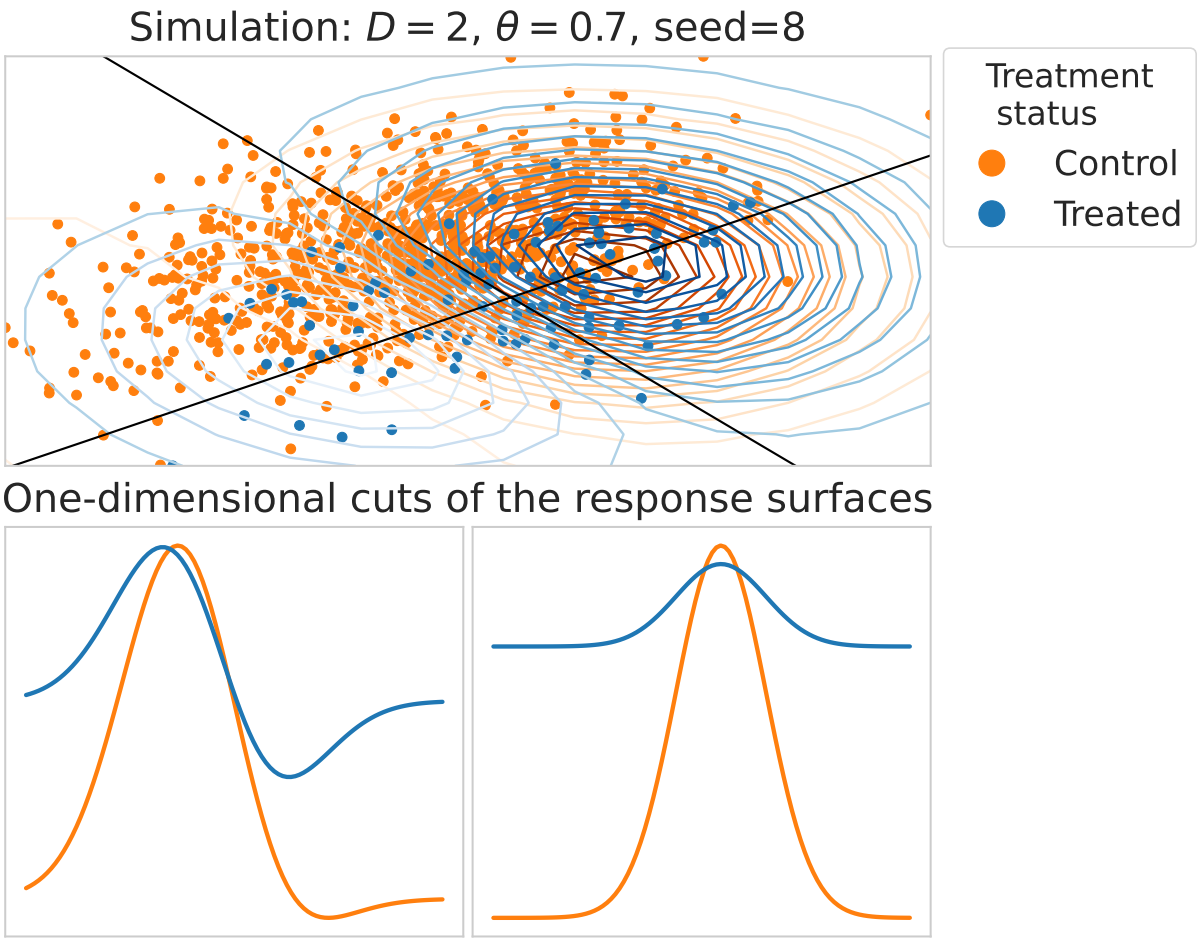


Figure S6: Example of the simulation setup in the input space with two knots –*ie.* basis functions. The top panel shows the observations in feature space, while the bottom panel displays the two response surfaces on a 1D cut along the black lines drawn on the top panel.

$D = 55$ 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 [76]: Starting from data from the Linked Births and Infant Deaths Database (LBIDD)[44] 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 we evaluate only feasible metrics, which do not require this nuisance parameter. We used all 432 datasets[7] of size $N = 5\,000$.

Twins [43]: It is an augmentation of real data on twin births and mortality rates[2]. There are $N = 11\,984$

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

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

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[15] 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:

$$\mathbf{t}_i \mid \mathbf{x}_i, \mathbf{z}_i \sim \text{Bern}\left(\sigma\left(w_o^\top \mathbf{x} + w_h(\mathbf{z}/10 - 0.1)\right)\right) \tag{18}$$
$$\text{with } w_o \sim \mathcal{N}(0, 0.1 \cdot I),\ w_h \sim \mathcal{N}(5, 0.1)$$

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.

G.3 Model selection procedures

Nuisances estimation The nuisances are estimated with a stacked regressor inspired by the Super Learner framework,[39]). 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 S2. All implementations come from scikit-learn[54]. As extreme inverse propensity weights induce high variance, we use clipping[84,32] 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 HistGradientBoostingRegressor learning rate: [0 HistGradientBoostingRegressor max leaf nodes
Treatment, e	StackedClassifier (HistgradientBoostingClassifier, LogisticRegression)	LogisticRegression C: [0.0001, 0.001, 0.01, 0.1, HistGradientBoostingClassifier learning rate: [0 HistGradientBoostingClassifier max leaf nodes:

Table S2: Hyper-parameters grid used for nuisance models

G.4 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

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

discordant pairs normalized by the total number of pairs. It takes values in the $[-1, 1]$ range.

$$\kappa = \frac{(\text{number of concordant pairs }) - (\text{number of discordant pairs})}{(\text{number of pairs})} \tag{19}$$

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 S7 - 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 S8 - Results measured in absolute Kendall's

Figure S9 - 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.

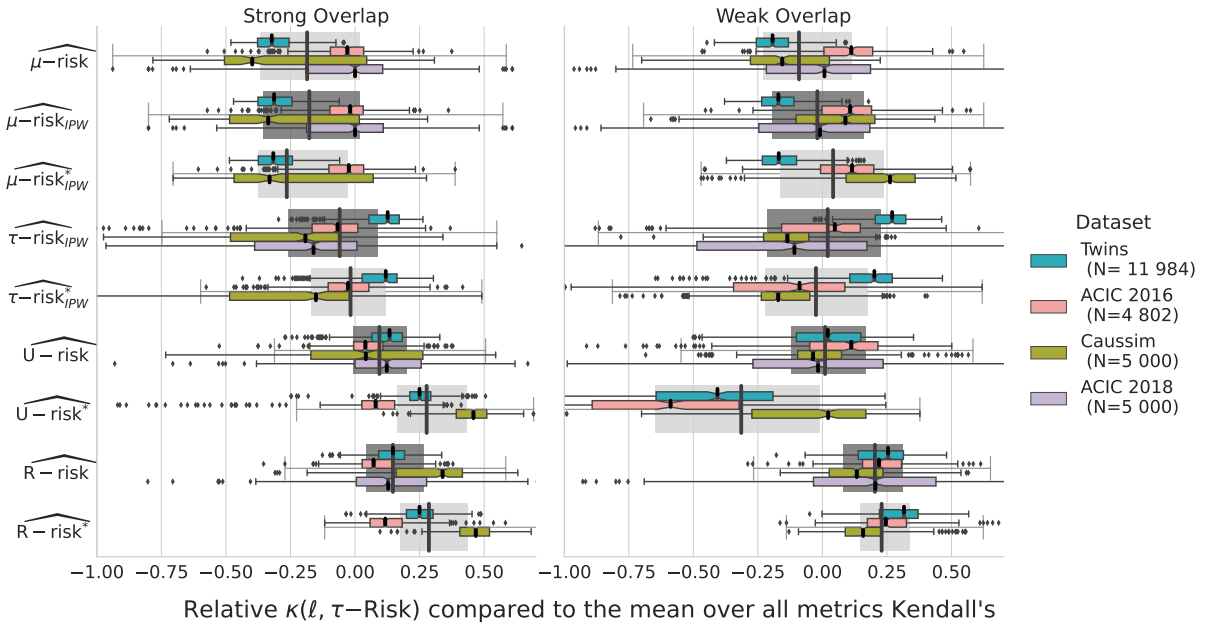


Figure S7: The R-risk is the best metric: Relative Kendall's τ agreement with τ-risk. Strong and Weak overlap correspond to the first and last tertiles of the overlap distribution measured with Normalized Total Variation eq. 16.

		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 S3: 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

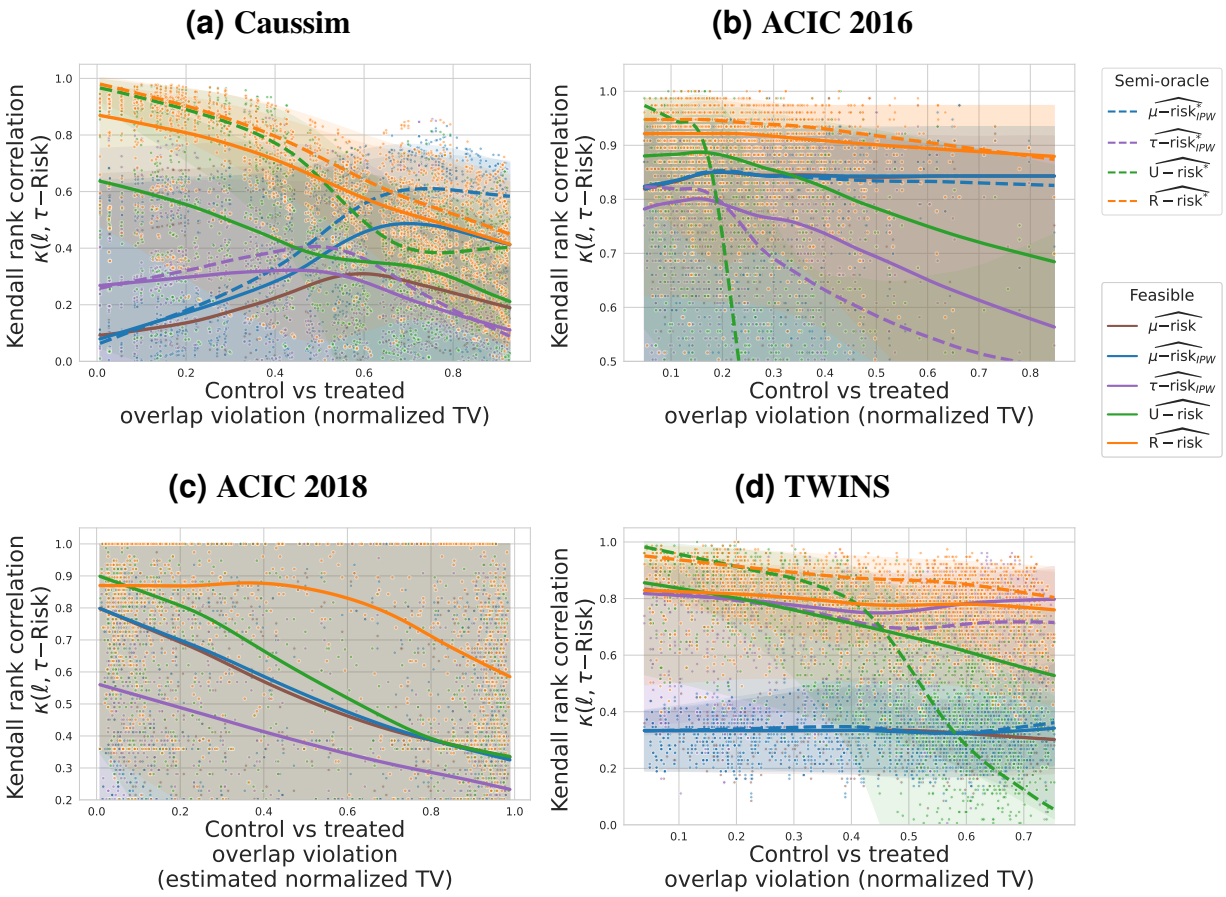


Figure S8: Agreement with τ-risk ranking of methods function of overlap violation. The lines represent medians, estimated with a lowess. The transparent bands denote the 5% and 95% confidence intervals.

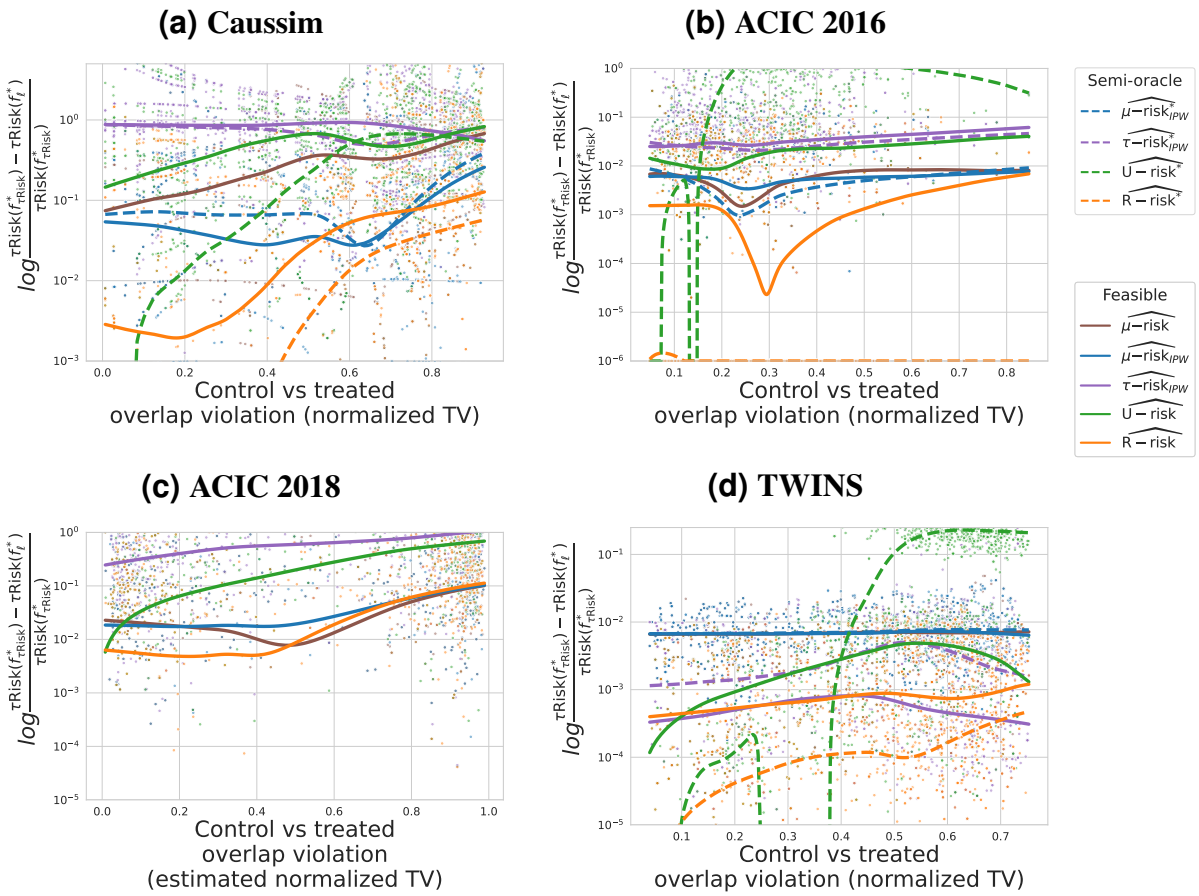


Figure S9: Metric performances by normalized tau-risk distance to the best method selected with τ-risk. All nuisances are learned with the same estimator stacking gradient boosting and ridge regression. Doted and plain lines corresponds to 60% lowess quantile estimates. This choice of quantile allows to see better the oracle metrics lines for which outliers with a value of 0 distord the curves.

Figure S10 - 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 S11 Low population overlap hinders model selection for all metrics

Figure S12 - 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 S13 - Flexible models are performant in recovering nuisances even in linear setups

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 S14 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$_{IPW}$.

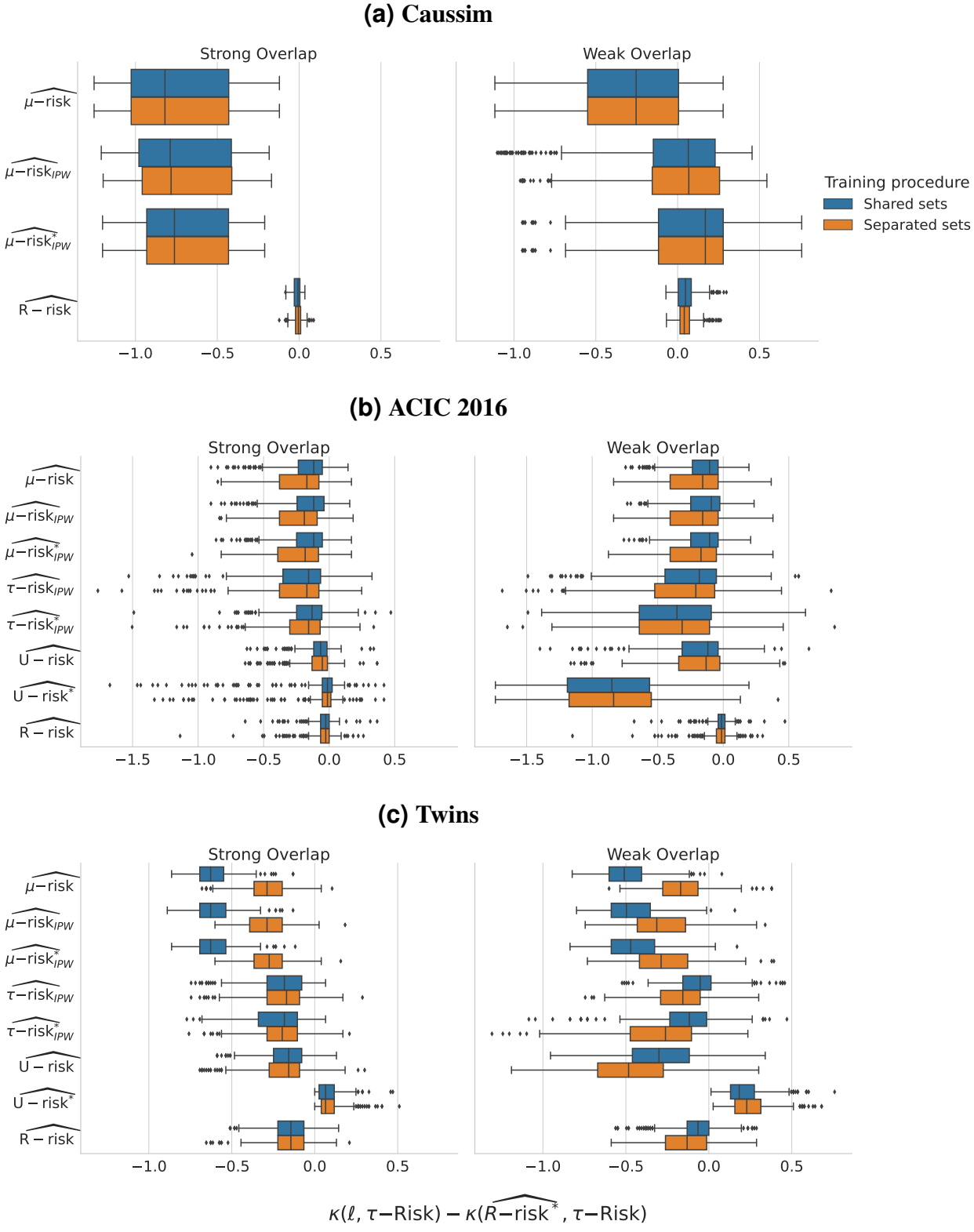


Figure S10: Results are similar between the Shared nuisances/candidate set and the Separated nuisances set procedure. The experience has not been run on the full metrics for Caussim due to computation costs.

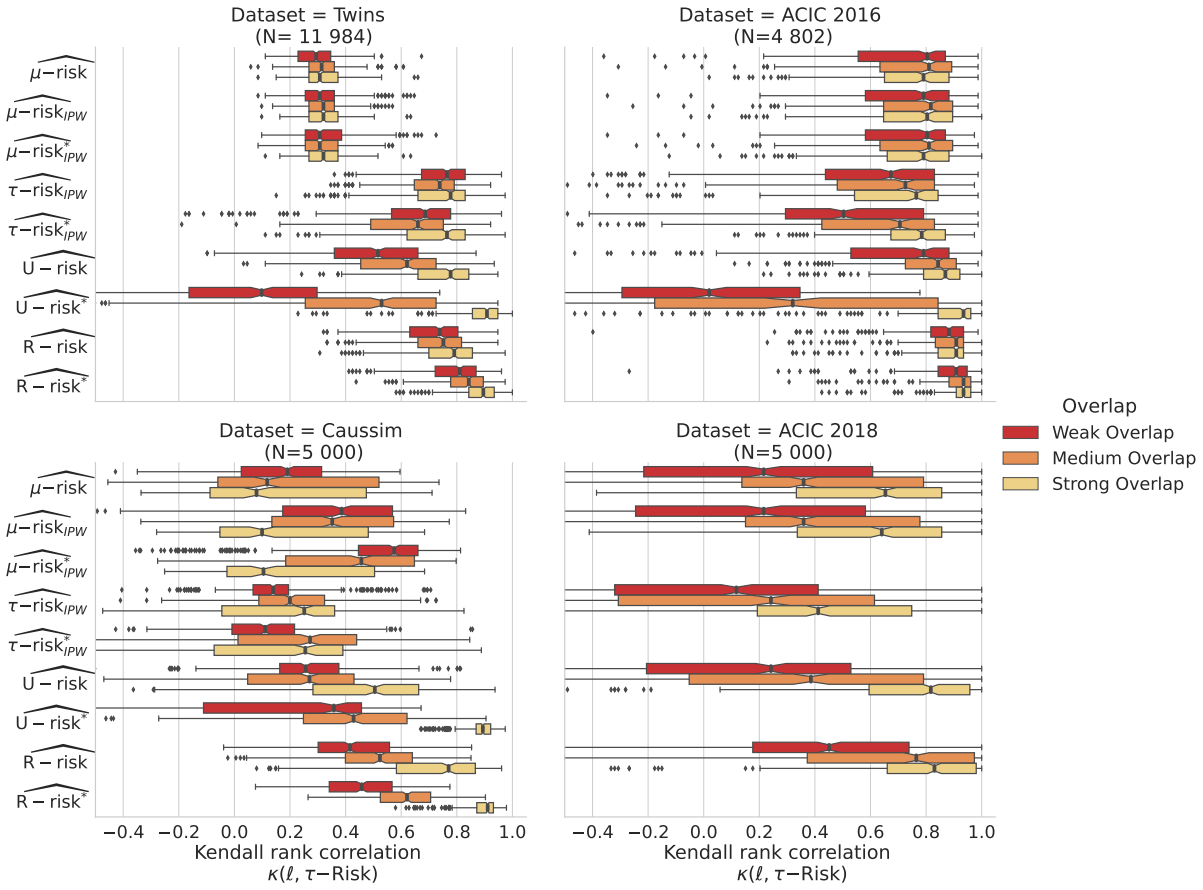


Figure S11: Low population overlap hinders causal model selection for all metrics: Kendall's τ agreement with τ-risk. Strong, medium and Weak overlap correspond to the tertiles of the overlap distribution measured with Normalized Total Varation eq. 16.

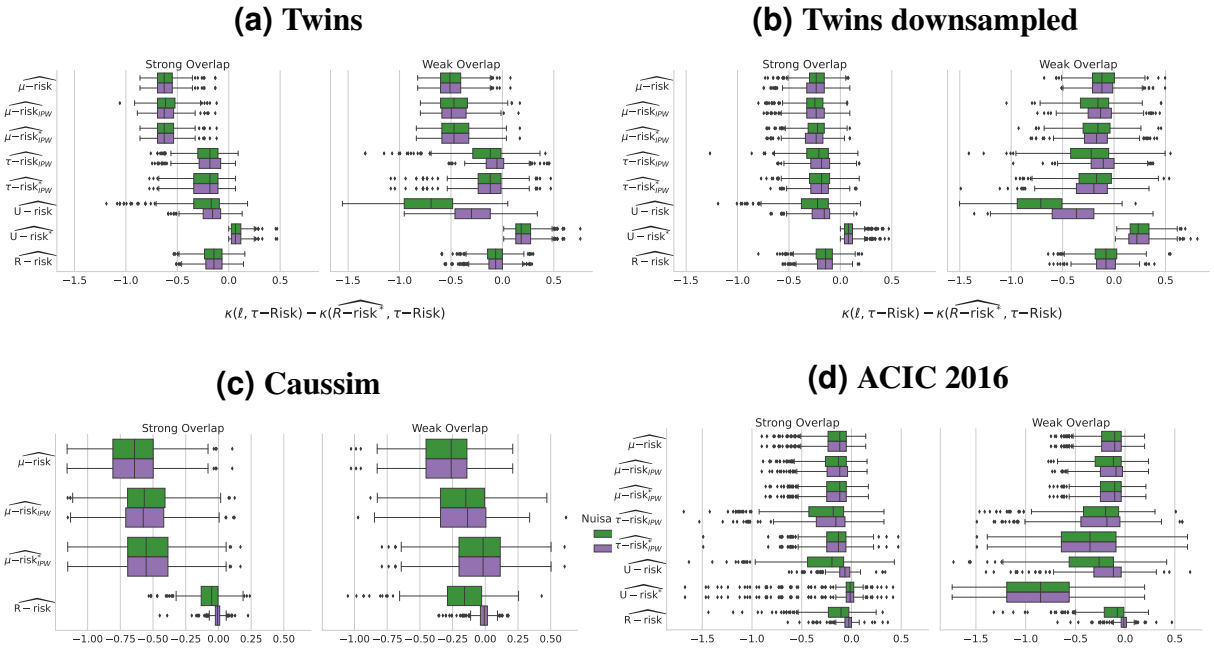


Figure S12: Learning the nuisances with stacked models (linear and gradient boosting) is important for successful model selection with R-risk. For Twins dataset, there is no improvement for stacked models compared to linear models because of the linearity of the propensity model.

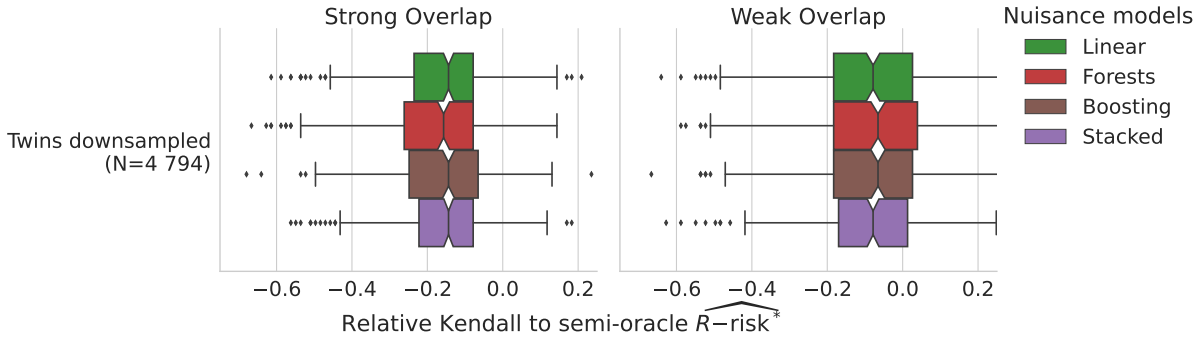
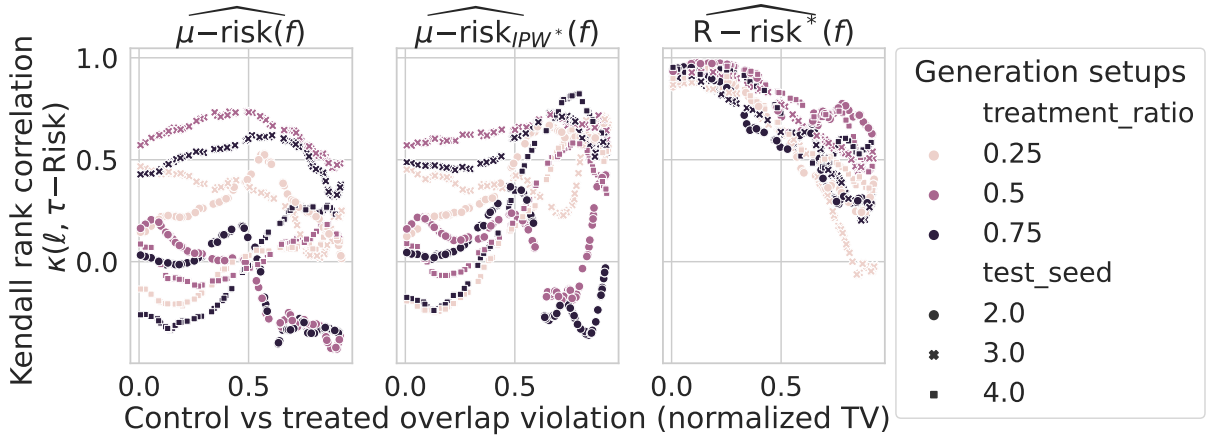


Figure S13: Flexible models are performant in recovering nuisances in the downsampled Twins dataset. The propensity score is linear in this setup, making it particularly challenging for flexible models compared to linear methods.

Figure S14: Kendall correlation coefficients for each causal metric. Each (color, shape) pair indicates a different (treatment ratio, seed) of the generation process.



H Data split choices

H.1 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 H.2).

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 S15 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.

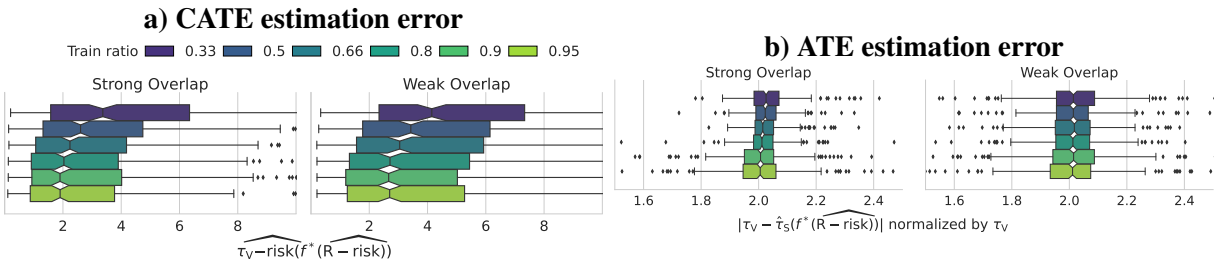


Figure S15: a) For CATE, a train/test ratio of 0.9/0.1 appears a good trade-off. b) For ATE, there is a small signal pointing also to 0.9/0.1 (K=10). for ATE. Experiences on 10 replications of all 78 instances of the ACIC 2016 data.

H.2 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[52]. Yet, it has been implemented with K=5 in causallib[75] or K=3 in econML[8]. Likewise, for ATE estimation, Chernozhukov et al. [14] 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: Loiseau et al. [42] use K=3, Gao et al. [21] use K=2. In the econML implementation, K is set to 3[8]. Naimi et al. [48] 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[24].

Causal model selection has been much less discussed. The only study that we are aware of, Schuler et al. [70], 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.