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

Matthieu Doutreligne, Gaël Varoquaux

How to select predictive models for decision making or causal inference

Matthieu Doutreligne[a,b,c], Gaël Varoquaux[a]

[a] *Inria Saclay, Soda team, Palaiseau, France*
[b] *Mission Data, Haute Autorité de Santé, Saint-Denis, France*
[c] *Corresponding author: matthieu.doutreligne@inria.fr*

Abstract

As predictive models –*eg* from machine learning– give likely outcomes, they may be used to support decision making, choosing whether to treat or not to treat. Such reasoning on the effect of an intervention is a causal-inference task. A plethora of predictive models are available for health data, from medical images to patient records. But, in a given situation, which of these models yield the most valid causal estimates? Here we highlight that classic machine-learning model selection does not select the best models for causal inference. Indeed, causal model selection should control both outcomes for each individual, treated or not treated, whereas only one outcome is observed. Theoretically, simple risks used in machine learning do not control causal effects when treated and non-treated population differ too much. More elaborate risks use "nuisance" re-weighting to approximate the causal error on the observed data. But does estimating these nuisance adds noise to model selection? Drawing from an extensive empirical study, we outline an efficient causal model-selection procedure. To select the best predictive model to guide decisions: use the so-called R-risk, use flexible estimators to compute the nuisance models on the train set, and split out 10% of the data to compute risks.

Keywords: Model Selection, Treatment Effect, G-formula, Observational Study, Machine Learning,

1. Introduction

1.1. Extending prediction to prescription needs causality

Progress in machine learning brings predictive models to new health data [10, 66], with the promise of precision medicine. Automated analysis of medical images is increasingly accurate, *eg* for brain images, [41, 97] or mammography [95, 81, 55]. New prognostic models leverage routinely-collected patient records [52]: predicting heart failure from claims [19], suicide attempts from questionnaires [84]... Clinical notes contain much prognostic information but require text modeling [33, 93, 86]. Data may be difficult to control and model, but the accuracy of the prediction can be verified on left-out data [3, 63, 91]. Given a model predicting a health outcome, precision medicine would like it to guide decisions: will an individual benefit from an intervention such as surgery [22]? Contrasting predictions with and without the treatment gives an answer, but statistical validity requires causal inference [85, 12].

There is a rich causal-inference statistical literature, which bridges to predictive modeling via so-called *outcome models*, or G-computation, G-formula [68], Q-model [85], conditional mean regression [94]. A central challenge of inference of treatment effects is that of confounding: spurious associations between treatment allocation and baseline health, *eg* only prescribing a drug to mild cases [29, 90]. Controlled allocation of treatment, as in randomized controlled trials (RCTs), alleviate this concern. Yet most machine-learning models are trained on *observational* data, close to real-word practice [11, 30] but challenging for causal inference. Causal inference has been central to epidemiology, typically with methods that model treatment assignment [7, 25], based on propensity scores [71]. Recent empirical results [94, 20] show benefits of outcome modeling to estimate average treatment effects. Maybe a greater benefit is that these methods naturally go beyond average effects, estimating individualized or conditional average treatment effects (CATE), central to precision medicine. For this purpose, such methods are also invaluable on randomized trials [88, 46, 32].

There is an explosion of outcome modeling or machine learning methods. For instance, many deep-learning methods have been developed for medical image analysis [79, 51]. Even outcome-modeling methods specifically designed for causal inference are numerous: Bayesian Additive Regression Trees [31], Targeted Maximum Likelihood Estimation [45, 77], causal boosting [64], causal multivariate adaptive regression splines [64], random forests [92, 5], Meta-learners [42], R-learners [57], Doubly robust estimation [15]... 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. For instance, Figure shows large variations obtained across different outcome estimators on semi-synthetic datasets [20]. 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

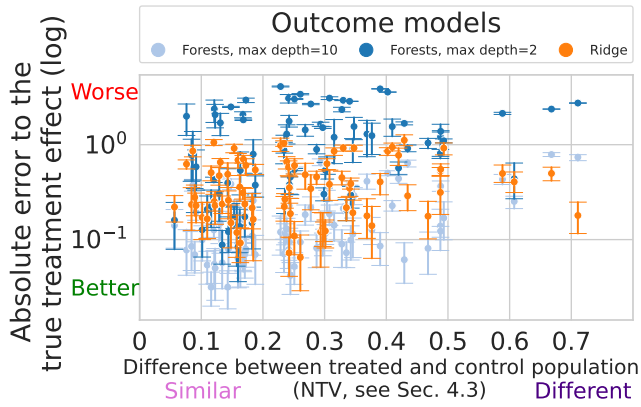


Figure 1: **Different outcome models lead to different estimation errors on the Average Treatment Effects**, on 77 classic simulations with known true causal effect [20]. The different models are ridge regression and random forests with different hyper-parameters (details Appendix A). The different configurations are plotted as a function of increasing difference between treated and untreated population –see subsection 4.3. There is no systematic best performer; data-driven model selection is important.

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.

Standard practices to select models in predictive settings rely on the error on the outcome [63, 91]. However, as we will see, these practices may not pick the best models for causal inference, as they can be misled by inhomogeneities due to treatment allocation. Given complex, potentially noisy, data, which model is to be most trusted to yield valid causal estimates? As no single learner performs best on all data sets, there is a pressing need for clear guidelines to select outcome models for causal inference.

Objectives and structure of the paper. In this paper, we study *model selection procedures* in practical settings: *finite samples* settings and without *well-specification* assumption. Asymptotic causal-inference theory calls for complex risks, but a practical question is whether model-selection procedures, that rely on data split, can estimate these risks reliably enough. Indeed, they 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 and briefly review prior art. Then, Section 2 sets causal model selection in the *potential outcome* framework and details the causal risks and model-selection procedure. Section 3 gives theoretical results. Section 4 details a thorough empirical study, covering many different settings. Finally, Section 5 discusses the findings. Results outline how to best select outcome models for causal inference with an adapted cross-validation to estimate the so-called R-risk. This risk compensates for systematic differences between treated and non-treated individuals using two *nuisance* models, themselves estimated from data and thus imperfect; yet these imperfections do not undermine the R-risk.

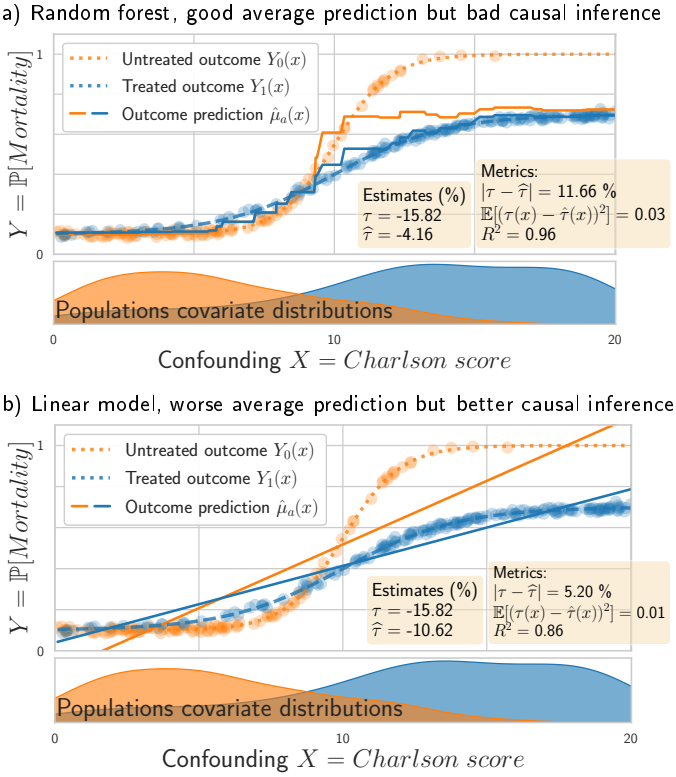


Figure 2: **Illustration**: a) a random-forest estimator with high performance for standard prediction (high R^2) but that yields poor causal estimates (large error between true effect τ and estimated $\hat{\tau}$), b) a linear estimator with smaller prediction performance leading to better causal estimation.

Selecting the estimator with the smallest error to the individual treatment effect $\mathbb{E}[(\tau(x) - \hat{\tau}(x))^2]$ –the τ-risk, def. 1 – 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.

1.2. 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 –as detailed in subsection 2.1. 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. However, this predictor may not be the best model to reason about causal effects of an intervention, as we illustrate below.

Figure 2 gives a toy example: the probability Y of an undesirable outcome (*eg* death), a binary treatment $A \in \{0, 1\}$, and a covariate $X \in \mathbb{R}$ summarizing the patient health status (eg. the Charlson index [14]). We simulate a treatment beneficial (decreases Y) for patients with high Charlson scores (bad health status) but with little effect for patients in good condition (low Charlson scores).

Figure 2a 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 2b 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– 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 criterion is not suited to choosing a model among candidate estimators for causal inference.

1.3. 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 [64], where the oracle CATE is known, Schuler et al. [76] compare four causal risks, concluding that for CATE estimation the best model-selection risk is the so-called R-risk [57] –def. 6, 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. 2– and reweighted MSE –μ-risk$_{IPW}$ def. 3– 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 [20], 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 [21]. 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 [70], an IPW estimate [27], a doubly robust estimator [74], 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 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) [45, 77] or Double Machine Learning [15]. Here also, theories require asymptotic regimes and models to be *well-specified*.

By contrast, Johansson et al. [38] 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 [76].

2. Formal setting: causal inference and model selection

2.1. The Neyman-Rubin Potential Outcomes framework

Settings. The Neyman-Rubin Potential Outcomes framework [54, 35] enables statistical reasoning on causal treatment effects: Given an outcome $Y \in \mathbb{R}$ (eg. mortality risk or hospitalization length), function of a binary treatment $A \in \mathcal{A} = \{0, 1\}$ (eg. a medical procedure, a drug administration), 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

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

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

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

Causal assumptions. Assumptions are necessary for causal estimands to be identifiability in observational settings [73]. We assume the usual strong ignorability assumptions: *1) unconfoundedness* $\{Y(0), Y(1)\} \perp\!\!\!\perp A|X$, *2) strong overlap* ie. every patient has a strictly positive probability to receive each treatment, *3) consistency*, and *4) generalization* (detailed in Appendix B).

Estimating treatment effects with outcome models. Should we know the two expected outcomes for a given X, we could compute the difference between them, which gives the causal effect of the treatment. These two expected outcomes can be computed from the observed data: the consistency 3 and unconfoundedness 1 assumptions imply the equality of two different expectations:

$$\mathbb{E}_{Y(a)\sim\mathcal{D}^\star}[Y(a)|X=x] = \mathbb{E}_{Y\sim\mathcal{D}}[Y|X=x, A=a] \quad (1)$$

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. This equality is referred as the g-formula identification [67]. 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 [85]), targeting the ATE τ with outcome modeling:

$$\begin{aligned}\tau = &\quad \mathbb{E}_{Y\sim\mathcal{D}^\star}[Y(1)-Y(0)|X=x] \\ = &\quad \mathbb{E}_{Y\sim\mathcal{D}}[Y|A=1] - \mathbb{E}_{Y\sim\mathcal{D}}[Y|A=0] \end{aligned} \quad (2)$$

This equation builds on two quantities: the conditional expectancy of the outcome given the covariates and either treatment or no no treatment, 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 2 leads to an estimator of the ATE written:

$$\hat{\tau} = \frac{1}{n}\left(\sum_{i=1}^{n} \mu_1(x_i) - \mu_0(x_i)\right) \quad (3)$$

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.

Similarly to the ATE, at the individual level, the CATE:

$$\tau(x) = \mu_1(x) - \mu_0(x) \quad (4)$$

Robinson decomposition. The *R-decomposition* of the outcome model plays an important role, [69]: introducing two quantities, the conditional mean outcome and the probability to be treated (known as propensity score [71]):

$$\text{Conditional mean outcome} \qquad m(x) \overset{\text{def}}{=} \mathbb{E}_{Y\sim\mathcal{D}}[Y|X=x] \quad (5)$$

$$\text{Propensity score} \qquad e(x) \overset{\text{def}}{=} \mathbb{P}[A=1|X=x] \quad (6)$$

the outcome can be written

$$\text{R-decomposition} \quad y(a) = m(x) + \big(a - e(x)\big)\tau(x) + \varepsilon(x;a) \quad \text{with} \quad \mathbb{E}[\varepsilon(X;A)|X,A] = 0 \quad (7)$$

m and e are often called *nuisances* [15]; they are unknown.

2.2. Model-selection risks, oracle and feasible

Causal model selection. We formalize model selection for causal estimation. Thanks to the g-formula identification (Equation 1), 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 3 and 4), $\hat{\tau}_f$ and $\hat{\tau}_f(x)$. 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:

$$f^*_\ell = \underset{f\in\mathcal{F}}{\operatorname{argmin}}\, \ell(f, O) \quad (8)$$

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:

Definition 1 (τ-risk(f)). *also called PEHE [75, 31]:*

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

Given observed data from $p(X)$, the expectation is computed with a finite sum, as in eq. 3, 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 $(Y, X, A) \sim \mathcal{D}$.

Feasible error risks. *Feasible* risks are based on the prediction error of the outcome model and *observable* quantities.

All expectations below are on observed distribution: $(Y, X, A) \sim \mathcal{D}$.

Definition 2 (Factual μ-risk). *[78] This is the usual Mean Squared Error on the target y. It is what is typically meant by "generalization error" in supervised learning:*

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

We now detail risks that use the nuisances e –propensity score, def 6– and m –conditional mean outcome, def 5. 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.

Definition 3 (μ-risk$^\star_{IPW}$). *[43] 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 4 (τ-risk$^\star_{IPW}$). *[92] The CATE $\tau(x)$ can be estimated with a regression against inverse propensity weighted outcomes [4, 27, 92], the τ-risk$_{IPW}$.*

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

Table 1: Review of causal risks — The R-risk* is called τ-risk$_R$ in Schuler et al. [76].

Risk	Equation	Reference
$mse(\tau(X), \tau_f(X)) = \tau$-risk	$\mathbb{E}_{X\sim p(X)}[(\tau(X) - \hat{\tau}_f(X))^2]$	Eq. 1 [31]
$mse(Y, f(X)) = \mu$-risk	$\mathbb{E}_{(Y,X,A)\sim\mathcal{D}}\left[(Y - f(X;A))^2\right]$	Def. 2 [76]
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. 3 [43]
τ-risk$^\star_{IPW}$	$\mathbb{E}_{(Y,X,A)\sim\mathcal{D}}\left[\left(Y\left(\frac{A}{e(X)} - \frac{1-A}{1-e(X)}\right) - \hat{\tau}_f(X)\right)^2\right]$	Def. 4 [92]
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. 5 [57]
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. 6 [57]

Definition 5 (U-risk*). *[42, 57] Based on the Robinson decomposition –eq. 7, the U-learner uses the $A - e(X)$ term in the denominator. The derived risk is:*

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

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 6 (R-risk*). *[57, 76] 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. 7. These risks are summarized in Table 1.

2.3. Estimation and model selection procedure

Causal model selection (as in Equation 8) 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 3.

As extreme inverse propensity weights induce high variance, clipping can be useful for numerical stability [89, 36].

3. Theory: Links between feasible and oracle risks

We now relate two feasible risks, μ-risk$_{IPW}$ and the R-risk to the oracle τ-risk. Both results make explicit the role of overlap for the performances of causal risks.

Algorithm 1 Model selection procedure

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

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

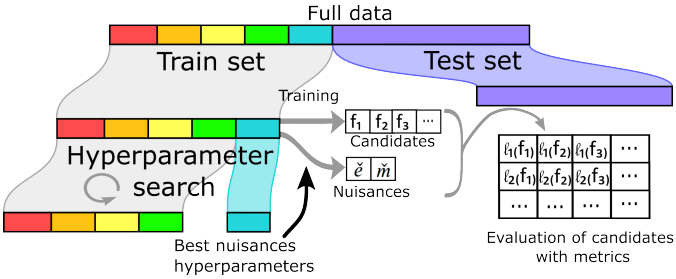


Figure 3: Estimation procedure for causal model selection.

These bounds depend on a specific form of residual that we now define: for each potential outcome, $a \in \{0; 1\}$, the variance conditionally on x is [78]:

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

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

3.1. Upper bound of τ-risk with μ-risk$_{IPW}$

Proposition 1 (Upper bound with μ-risk$_{IPW}$). *[38] Given an outcome model f, let a weighting function $w(x;a) = \frac{a}{e(x)} + \frac{1-a}{1-e(x)}$ as the Inverse Propensity Weight. Then, under overlap (assumption 2), we have:*

$$\tau\text{-}risk(f) \leq\ 2\,\mu\text{-}risk_{IPW}(w, f) - 2\left(\sigma_B^2(1) + \sigma_B^2(0)\right)$$

This result has been derived in previous work [38]. It links μ-risk$_{IPW}$ to the squared residuals of each population. For completeness, we provide the proof in Appendix C.

The upper-bound comes from the triangular inequality applied to the residuals of both populations. The two quantities are equal when the absolute residuals on treated and untreated populations are equal on the whole covariate space: $\forall x \in \mathcal{X}, |\mu_1(x) - f(x,1)| = |\mu_0(x) - f(x,0)|$. The main difference between the oracle τ-risk and the reweighted mean squared error, μ-risk$_{IPW}$, comes from heterogeneous residuals between populations. This bound shows that minimizing the μ-risk$_{IPW}$ helps to minimize the τ-risk, which leads to interesting optimization procedures [38]. However, there is no guarantee that this bound is tight, which makes it fragile for model selection.

Assuming strict overlap (probability of all individuals being treated or not bounded away from 0 and 1 by η, Appendix B), the above bound simplifies into a looser one involving the usual mean squared error: τ-risk$(f) \leq \frac{2}{\eta}\,\mu$-risk$(f) - 2\left(\sigma_B^2(1) + \sigma_B^2(0)\right)$. For weak overlap (propensity scores not bounded far from 0 or 1), this bound is very loose (as shown in Figure 2) and is not appropriate to discriminate between models with close performances.

3.2. Reformulation of the R-risk as reweighted τ-risk

We now derive a novel rewriting of the R-risk, making explicit its link with the oracle τ-risk.

Proposition 2 (R-risk as reweighted τ-risk). *Given an outcome model f, its R-risk appears as weighted version of its τ-risk (Proof in Appendix C.2):*

$$R\text{-risk}^*(f) = \int_x e(x)\big(1-e(x)\big)\big(\tau(x) - \tau_f(x)\big)^2 p(x)dx + \tilde{\sigma}_B^2(1) + \tilde{\sigma}_B^2(0) \tag{9}$$

The R-risk targets the oracle at the cost of an overlap re-weighting and the addition of the reweighted Bayes residuals, which are independent of f. In good overlap regions the weights $e(x)\big(1-e(x)\big)$ are close to $\frac{1}{4}$, hence the R-risk is close to the desired gold-standard τ-risk. On the contrary, for units with extreme overlap violation, these weights go down to zero with the propensity score.

3.3. 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 noted in Section 3.1 and shown experimentally in Schuler et al. [76] and our results below.

Concerning the R-risk, replacing $e(x)$ by its randomized value p_A in Proposition 2 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 [76] 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:

$$\tau\text{-risk}(\mu) = \mathbb{E}_{X\sim p(X)}[(\tau(X) - \tau_\mu(X))^2] = 0 \tag{10}$$

$$\begin{aligned}\mu\text{-risk}(\mu) &= \mathbb{E}_{(Y,X,A)\sim p(Y;X;A)}\big[\big(Y - \mu_A(X)\big)^2\big] \\ &= \int_{\mathcal{X},\mathcal{A}} \varepsilon(x,a)^2 p(a \mid x)\, p(x)\, dx\, da \leq \sigma_B^2(0) + \sigma_B^2(1)\end{aligned} \tag{11}$$

$$\mu\text{-risk}_{IPW}(\mu) = \sigma_B^2(0) + \sigma_B^2(1) \quad \text{from Lemma 1} \tag{12}$$

$$R\text{-risk}(\mu) = \tilde{\sigma}_B^2(0) + \tilde{\sigma}_B^2(1) \leq \sigma_B^2(0) + \sigma_B^2(1) \quad \text{from Proposition 2} \tag{13}$$

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.

4. 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].

4.1. 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 [76, 1], we use many generative processes, to reach more general conclusions (as discussed in E.21).

We generate the response functions using random bases. Basis extension methods are common in biostatistics, *eg* functional regression with splines [34, 61]. By allowing the function to vary at specific knots, they give flexible

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

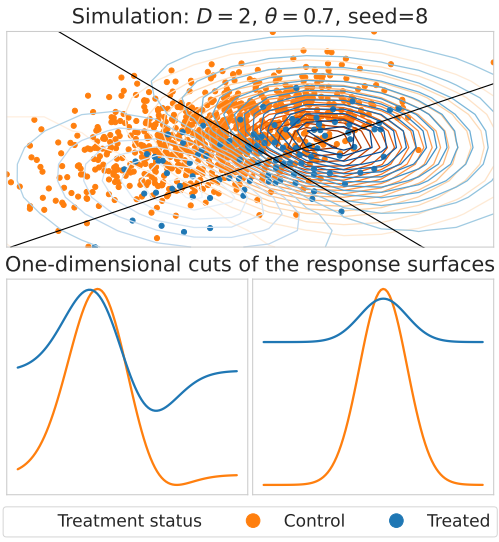


Figure 4: Example of the simulation setup in the input space with two knots –*ie.*basis functions. The top panel shows the observations in feature space, while the bottom panel displays the two response surfaces on a 1D cut along the black lines drawn on the top panel.

non-linear models. We use random approximation of Radial Basis Function (RBF) kernels [65] to generate the response functions. RBF use the same process as polynomial splines but replace polynomial by Gaussian kernels. Unlike polynomials, Gaussian kernels have decreasing influences in the input space. This avoids unrealistic divergences of the response surfaces at the ends of the feature space.

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 section 3.1). Finally, we generate observed outcomes adding Gaussian noise. We generate 1 000 of such datasets, with uniformly random overlap parameters. Details in Appendix E.1.

Family of candidate estimators. We test model selection on a family of candidate estimators that approximate imperfectly the data-generating process. To build such an estimator, we first use a RBF expansion similar to the one used for data generation. We choose two random knots and apply a transformation of 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 mode; we use common nomenclature [42, 80] 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, 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 treat and the non treated, 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.

We do not include more elaborated meta-learners such as R-learner [57] or X-learner [42]. Our goal is not to have the best possible learner but to have a variety of sub-optimal learners to compare the different causal metrics. For the same reason, we did not include more powerful outcome models such as random forests or boosting trees.

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.

4.2. Semi-simulated datasets

Datasets. We also use classic benchmarks of the causal-inference literature, semi-simulated data adding a known synthetic causal effect to real –non synthetic– covariate:

ACIC 2016 [20]: The dataset is based on the Collaborative Perinatal Project [58], a RCT studying infants' developmental disorders. 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 [2]. We used 10 different seeds for each setup, totaling 770 dataset instances.

ACIC 2018 [83]: Starting from data from the Linked Births and Infant Deaths Database (LBIDD) [49] 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[3] of size $N = 5\,000$.

Twins [48]: It is an augmentation of real data on twin births and mortality rates [2]. There are $N = 11\,984$

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

[3]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[4], The outcome is the mortality and the treatment is the weight of the heavier twin at birth. This is a "true" counterfactual dataset [17] 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:

$$\mathbf{t}_i \mid \mathbf{x}_i, \mathbf{z}_i \sim \text{Bern}\left(\sigma\left(w_o^\top \mathbf{x} + w_h(\mathbf{z}/10 - 0.1)\right)\right) \quad (14)$$
$$\text{with } w_o \sim \mathcal{N}(0, 0.1 \cdot I),\ w_h \sim \mathcal{N}(5, 0.1)$$

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.

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 [5] 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 inspiration from the TMLE literature that uses combination of flexible machine learning methods [77], we use as models for the nuisances $\check{e}$ (respectively $\check{m}$) a form of meta-learner: a stacked estimator of ridge and boosting classifiers (respectively regressions). We select hyper-parameters with randomized search on a validation set $\mathcal{V}$ and keep them fix for model selection (Appendix E.2 lists hyperparameters). As extreme inverse propensity weights induce high variance, we use clipping [89, 36] to bound $min(\check{e}, 1 - \check{e})$ away from 0 with a fixed $\eta = 10^{-10}$, ensuring strict overlap for numerical stability.

4.3. Measuring overlap between treated and non treated

Good overlap between treated and control population is crucial for causal inference as it is required by the positivity assumption 2. It is often assessed by comparing visually population distributions (as in Figure 2) 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 $\mathbb{P}(X|A = 1)$ [21, 38]. We introduce the Normalized Total Variation (NTV), a divergence based on the sole propensity score (see Appendix D).

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

[5] Scikit-learn regressor, HistGradientBoostingRegressor, and classifier, HistGradientBoostingClassifier.

4.4. Results: factors driving good model selection

The R-risk is the best metric. Each metric ranks differently the candidate models. Figure 5 shows the agreement between the ideal ranking of methods given the oracle τ-risk and the different feasible causal metrics. We measure this agreement with a relative[6] Kendall tau κ (eq. E.2) [39]. Given the importance of overlap in how well metrics approximate the oracle τ-risk (Appendix C.1), 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 [4, 27]–leads to 100% excess risk (Figure E.16). Across datasets, the R-risk consistently decreases the risk compared to the μ-risk: 0.1%

[6] 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 $\kappa(\ell, \tau\text{−risk}) = \kappa(\ell, \tau\text{−risk}) - mean_\ell\big(\kappa(\ell, \tau\text{−risk})\big)$

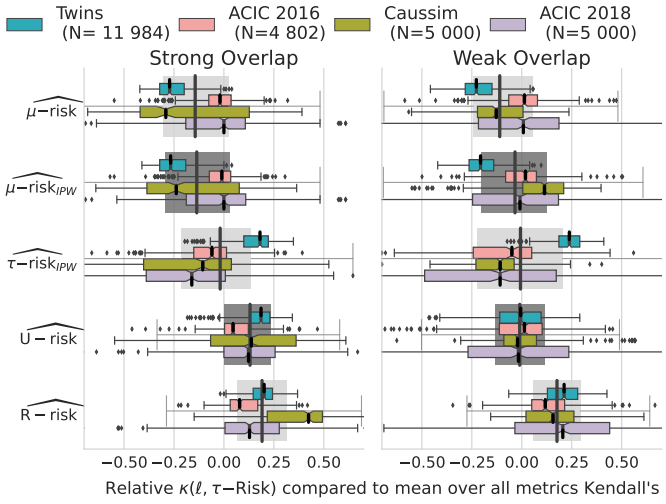


Figure 5: **The R-risk is the best metric**: Relative Kendall's τ agreement with τ-risk. Strong and Weak overlap correspond to the first and last tertiles of the overlap distribution measured with Normalized Total Variation eq. D.1. Appendix E.3 presents the same results by adding semi-oracle risks in Figure E.14, measured with absolute Kendall's in Figure E.15 and with τ−risk gains in Figure E.16. Table E.4 displays the median and IQR for the relative Kendall's results.

compared to 1% on ACIC2016, 1% compared to 20% on ACIC2018, and 0.05% compared to 1% on Twins.

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 6). The absolute Kendall's coefficient correlation with τ-risk drops from values around 0.9 (excellent agreement with oracle selection) to 0.6 on both Caussim and ACIC 2018 (Appendix E.3).

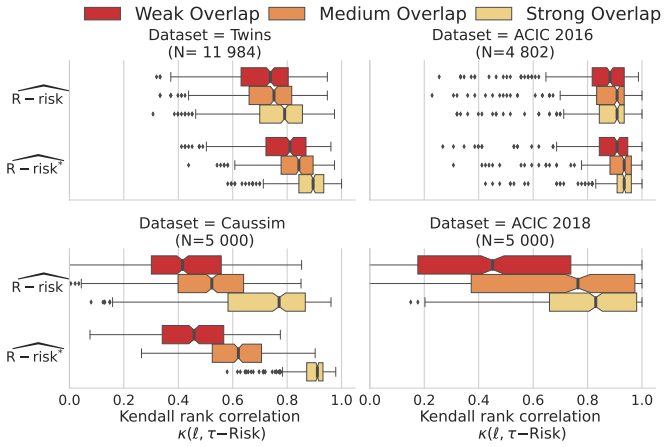


Figure 6: **Model selection is harder for low population overlap**: Kendall's τ agreement with τ-risk. Strong, medium and Weak overlap are the tertiles of the overlap measured with NTV eq. D.1. Appendix E.3 presents results for all metrics in Figure E.18 in absolute Kendall's and continuous overlap values in Figure E.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 [57]. In theory, these correlations can bias model selection and, strictly speaking, push to split out a third separated data set –a "nuisance set"– to fit the nuisance models. The drawback is that it depletes the data available for model estimation and selection. However, Figure 7 shows no substantial difference between a procedure with a separated nuisance set and the simpler shared nuisance-candidate set procedure.

Stacked models are good overall estimators of nuisances. For every risk, the oracle version recovers better the best estimator. However, 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 8 compares causal model selection estimating nuisances with stacked estimators or linear model. It comprises the Twins data, where the true propensity model is linear, and a downsampled version of this data, to study

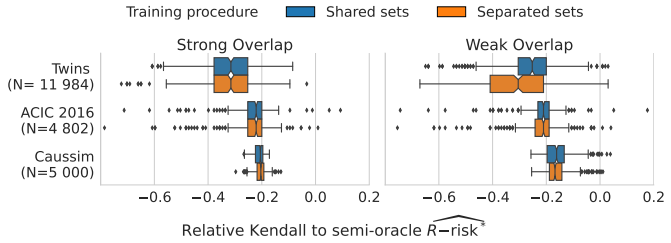


Figure 7: **Nuisances can be estimated on the same data as outcome models**: Results for the R-risk are similar between the shared nuisances/candidate set and the separated nuisances set procedures. Figure E.17 details results for all metrics.

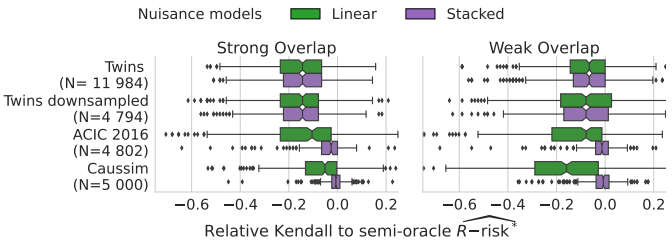


Figure 8: **Stacked models are good overall estimators of the nuisances**: Results are shown only for the R-risk; Figure E.19 details every metrics. For Twins, where the true propensity model is linear, stacked and linear estimations of the nuisances performs equivalently, even for a downsampled version (N=4794).

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

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 Appendix F).

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

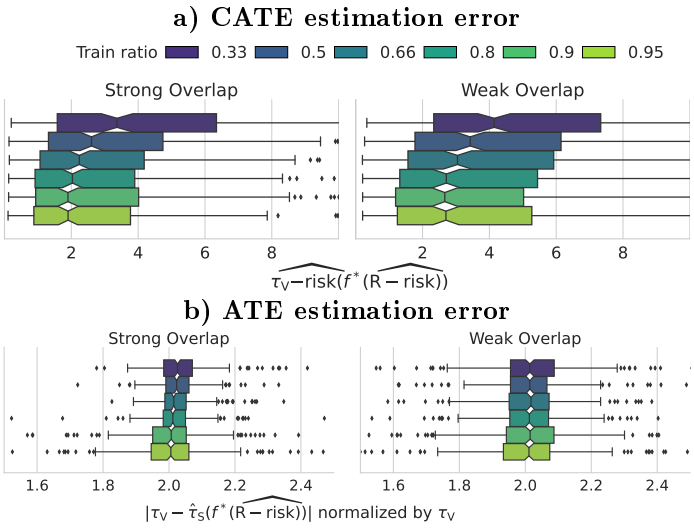


Figure 9: **a) For CATE, a train/test ratio of 0.9/0.1 appears a good trade-off.** b) For ATE, there is a small signal pointing also to 0.9/0.1 (K=10). for ATE. Experiences on 10 replications of all 78 instances of the ACIC 2016 data.

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 9 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.

5. Discussion and conclusion

Predictive models are increasingly used to reason about causal 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 (estimating the so-called μ-risk). Rather, selecting the best outcome model according to the R-risk (eq. 6) leads to more valid causal estimates.

Nuisance models: more gain than pain. 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, though our results show that these can be learned on the same set of data as the outcome model evaluated.

The nuisance models must be well estimated (Figure 8). However these models are easier to select and control than a causally-valid outcome model, as they are associated to errors on observed distributions. Our results show that using for nuisance models a flexible stacking-based family of estimator suffices for good model selection. In fact, a feasible R-risk –where the nuisances are estimated– performs almost as well as an oracle R-risk –where the nuisances are known. This may be explained by results that suggest that estimation errors on both nuisances partly compensate out in the R-risk[18, 40, 57, 15, 98, 53].

Note that propensity score models must be selected to estimate the individual posterior probability. For this, we used the Brier score, is minimized by the true individual probability. An easy mistake is to use calibration errors popular in machine learning [62, 96, 56, 50] as these select not for the individual posterior probability but for an aggregate error rate [60].

Extension to binary outcomes. While we focused on continuous outcomes, in medicine, the target outcome is often a categorical variable such as mortality status or diagnosis. In this case, it may be interesting to focus on other estimands than the Average Treatment Effect $\mathbb{E}[Y(1)] - \mathbb{E}[Y(0)]$, for instance the relative risk $\frac{\mathbb{P}(Y(1)=1)}{\mathbb{P}(Y(0)=1)}$ or the odd ratio, $\frac{\mathbb{P}(Y(1)=1)/[1-\mathbb{P}(Y(1)=1)]}{\mathbb{P}(Y(0)=1)/[1-\mathbb{P}(Y(0)=1]}$ are often used [8]. While the odds ratio is natural for case-control studies [72], other measures can reduce heterogeneity [16]. In the log domain, the ratios are written as a difference, the framework studied here (subsection 2.1) can directly apply. In particular, the log odds ratio is estimated by the common cross-entropy loss (or log loss) as in logistic regression.

More R-risk to select models driving decisions. Prediction models have flourished because their predictions can be easily demonstrated and validated on left-out data. But they require more careful validation for decision making, using a metric accounting for the putative intervention, the R-risk. Even when treated and untreated population differ little, as in RCTs, the R-risk brings a sizeable benefit. To facilitate better model selection, we provide Python code[7]. Using the R-risk does make evaluation more complicated not only because the procedure is more involved, but also because each intervention requires a dedicated evaluation. However, such off-policy evaluation remains much less costly than the recommended good practice of impact evaluation testing the ability of a prediction model to actually guide patient health [28]. Also, the 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, or language models.

Acknowledgments

We acknowledge fruitful discussions with Bénédicte Colnet. GV acknowledges partial funding from Projet Intercept-T2D (HORIZON-HLTH-2022-STAYHLTH-02-01).

Financial disclosure

None reported.

[7] https://github.com/soda-inria/causal_model_selection

Conflict of interest

The authors declare no potential conflict of interests.

References

[1] Alaa, A., Schaar, M.V.D., . Validating causal inference models via influence functions. International Conference on Machine Learning , 191–201.

[2] Almond, D., Chay, K.Y., Lee, D.S., 2005. The Costs of Low Birth Weight. The Quarterly Journal of Economics 120.

[3] Altman, D.G., Vergouwe, Y., Royston, P., Moons, K.G., 2009. Prognosis and prognostic research: validating a prognostic model. Bmj 338.

[4] Athey, S., Imbens, G., 2016. Recursive partitioning for heterogeneous causal effects. Proceedings of the National Academy of Sciences 113, 7353–7360.

[5] Athey, S., Tibshirani, J., Wager, S., . Generalized random forests. Annals of Statistics 47, 1148–1178.

[6] Austin, P.C., 2011. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behavioral Research , 399–424.

[7] Austin, P.C., Stuart, E.A., . 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, 3661–3679.

[8] Austin, P.C., Stuart, E.A., 2017. Estimating the effect of treatment on binary outcomes using full matching on the propensity score. Statistical methods in medical research 26, 2505–2525.

[9] Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oka, P., Oprescu, M., Syrgkanis, V., . EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation. https://github.com/microsoft/EconML.

[10] Beam, A.L., Kohane, I.S., 2018. Big data and machine learning in health care. Jama 319, 1317–1318.

[11] Black, N., 1996. Why we need observational studies to evaluate the effectiveness of health care. Bmj 312, 1215–1218.

[12] Blakely, T., Lynch, J., Simons, K., Bentley, R., Rose, S., 2020. Reflection on modern methods: when worlds collide–prediction, machine learning and causal inference. International journal of epidemiology 49, 2058–2064.

[13] Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Mohammadi Sepahvand, N., Raff, E., Madan, K., Voleti, V., Ebrahimi Kahou, S., Michalski, V., Arbel, T., Pal, C., Varoquaux, G., Vincent, P., . Accounting for variance in machine learning benchmarks. Proceedings of Machine Learning and Systems 3, 747–769.

[14] Charlson, M.E., Pompei, P., Ales, K.L., MacKenzie, C., . A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation. Journal of Chronic Diseases 40, 373–383.

[15] Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., . Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal , 71.

[16] Colnet, B., Josse, J., Varoquaux, G., Scornet, E., 2023. Risk ratio, odds ratio, risk difference... which causal measure is easier to generalize? arXiv preprint arXiv:2303.16008 .

[17] Curth, A., Svensson, D., Weatherall, J., . Really doing great at estimating CATE? a critical look at ML benchmarking practices in treatment effect estimation. Neurips Process 2021 , 14.

[18] Daniel, R.M., 2018. Double Robustness. John Wiley & Sons, Ltd. pp. 1–14.

[19] Desai, R.J., Wang, S.V., Vaduganathan, M., Evers, T., Schneeweiss, S., 2020. 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, e1918962–e1918962.

[20] Dorie, V., Hill, J., Shalit, U., Scott, M., Cervone, D., 2019. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. Statistical Science 34, 43–68.

[21] D'Amour, A., Ding, P., Feller, A., Lei, L., Sekhon, J., 2021. Overlap in observational studies with high-dimensional covariates. Journal of Econometrics 221, 644–654.

[22] Fontana, M.A., Lyman, S., Sarker, G.K., Padgett, D.E., MacLean, C.H., 2019. Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty? Clinical orthopaedics and related research 477, 1267.

[23] Gao, Z., Hastie, T., Tibshirani, R., 2021. Assessment of heterogeneous treatment effect estimation accuracy via matching. Statistics in Medicine .

[24] Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A., . A kernel two-sample test. The Journal of Machine Learning Research 13, 723–773.

[25] Grose, E., Wilson, S., Barkun, J., Bertens, K., Martel, G., Balaa, F., Khalil, J.A., . Use of propensity score methodology in contemporary high-impact surgical literature. Journal of the American College of Surgeons 230, 101–112.e2.

[26] Gruber, S., van der Laan, M.J., 2012. tmle: An R package for targeted maximum likelihood estimation. Journal of Statistical Software 51, 1–35.

[27] Gutierrez, P., Gerardy, J.Y., . Causal inference and uplift modeling a review of the literature. Proceedings of The 3rd International Conference on Predictive Applications and APIs , 14.

[28] Hendriksen, J.M., Geersing, G.J., Moons, K.G., de Groot, J.A., 2013. Diagnostic and prognostic prediction models. Journal of Thrombosis and Haemostasis 11, 129–141.

[29] Hernán, M., Robins, J., . Causal Inference: What If. CRC Boca Raton, FL.

[30] Hernán, M.A., . Methods of public health research — strengthening causal inference from observational data. New England Journal of Medicine 385, 1345–1348.

[31] Hill, J.L., . Bayesian nonparametric modeling for causal inference. Journal of Computational and Graphical Statistics , 217–240.

[32] Hoogland, J., IntHout, J., Belias, M., Rovers, M.M., Riley, R.D., E. Harrell Jr, F., Moons, K.G., Debray, T.P., Reitsma, J.B., 2021. A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint. Statistics in medicine 40, 5961–5981.

[33] Horng, S., Sontag, D.A., Halpern, Y., Jernite, Y., Shapiro, N.I., Nathanson, L.A., 2017. Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning. PloS one 12, e0174708.

[34] Howe, C.J., Cole, S.R., Westreich, D.J., Greenland, S., Napravnik, S., Eron, J.J., . Splines for trend analysis and continuous confounder control 22, 874–875.

[35] Imbens, G.W., Rubin, D.B., . Causal inference in statistics, social, and biomedical sciences. Cambridge University Press.

[36] Ionides, E.L., . Truncated importance sampling. Journal of Computational and Graphical Statistics 17, 295–311.

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

[38] Johansson, F.D., Shalit, U., Kallus, N., Sontag, D., 2022. Generalization bounds and representation learning for estimation of potential outcomes and causal effects.

[39] Kendall, M.G., 1938. A new measure of rank correlation. Biometrika 30, 81–93.

[40] Kennedy, E.H., 2020. Optimal doubly robust estimation of heterogeneous causal effects. arXiv preprint arXiv:2004.14497 .

[41] Khojaste-Sarakhsi, M., Haghighi, S.S., Ghomi, S.F., Marchiori, E., 2022. Deep learning for alzheimer's disease diagnosis: A survey. Artificial Intelligence in Medicine , 102332.

[42] Künzel, S.R., Sekhon, J.S., Bickel, P.J., Yu, B., . Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences 116, 4156–4165.

[43] van der Laan, M.J., Laan, M., Robins, J., . Unified methods for censored longitudinal data and causality. Springer Science & Business Media.
[44] Laan, M.J.v.d., Polley, E.C., Hubbard, A.E., . Super learner 6.
[45] Laan, M.J.v.d., Rose, S., . Targeted Learning. Springer Series in Statistics.
[46] Lamont, A., Lyons, M.D., Jaki, T., Stuart, E., Feaster, D.J., Tharmaratnam, K., Oberski, D., Ishwaran, H., Wilson, D.K., Van Horn, M.L., 2018. Identification of predicted individual treatment effects in randomized clinical trials. Statistical methods in medical research 27, 142–157.
[47] Loiseau, N., Trichelair, P., He, M., Andreux, M., Zaslavskiy, M., Wainrib, G., Blum, M.G.B., 2022. External control arm analysis: an evaluation of propensity score approaches, G-computation, and doubly debiased machine learning. BMC Medical Research Methodology 22.
[48] Louizos, C., Shalit, U., Mooij, J., Sontag, D., Zemel, R., Welling, M., . Causal effect inference with deep latent-variable models. Advances in neural information processing systems .
[49] MacDorman, M.F., Atkinson, J.O., 1998. Infant mortality statistics from the linked birth/infant death data set–1995 period data. Monthly Vital Statistics Report 46, 1–22.
[50] Minderer, M., Djolonga, J., Romijnders, R., Hubis, F., Zhai, X., Houlsby, N., Tran, D., Lucic, M., 2021. Revisiting the Calibration of Modern Neural Networks. Advances in Neural Information Processing Systems 34, 15682–15694.
[51] Monshi, M.M.A., Poon, J., Chung, V., 2020. Deep learning in generating radiology reports: A survey. Artificial Intelligence in Medicine 106, 101878.
[52] Mooney, S.J., Pejaver, V., 2018. Big data in public health: terminology, machine learning, and privacy. Annual review of public health 39, 95.
[53] Naimi, A.I., Mishler, A.E., Kennedy, E.H., 2021. Challenges in obtaining valid causal effect estimates with machine learning algorithms. American Journal of Epidemiology .
[54] Naimi, A.I., Whitcomb, B.W., 2023. Defining and identifying average treatment effects. American Journal of Epidemiology .
[55] Nassif, A.B., Talib, M.A., Nasir, Q., Afadar, Y., Elgendy, O., 2022. Breast cancer detection using artificial intelligence techniques: A systematic literature review. Artificial Intelligence in Medicine , 102276.
[56] Niculescu-Mizil, A., Caruana, R., 2005. Predicting good probabilities with supervised learning, in: Proceedings of the 22nd international conference on Machine learning - ICML '05, ACM Press. pp. 625–632.
[57] Nie, X., Wager, S., . Quasi-oracle estimation of heterogeneous treatment effects. Biometrika 108, 299–319.
[58] Niswander, K.R., Stroke, U.S.N.I.o.N.D.a., . 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: A0bdVhlhDQkC.
[59] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., . Scikit-learn: Machine learning in python. Journal of Machine Learning Research 12, 2825–2830.
[60] Perez-Lebel, A., Morvan, M.L., Varoquaux, G., 2022. Beyond calibration: estimating the grouping loss of modern neural networks. arXiv preprint arXiv:2210.16315 .
[61] Perperoglou, A., Sauerbrei, W., Abrahamowicz, M., Schmid, M., . A review of spline function procedures in r 19, 46.
[62] Platt, J.C., Platt, J.C., 1999. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. Advances in Large Margin Classifiers , 61–74.
[63] Poldrack, R.A., Huckins, G., Varoquaux, G., 2020. Establishment of best practices for evidence for prediction: a review. JAMA psychiatry 77, 534–540.
[64] Powers, S., Qian, J., Jung, K., Schuler, A., Shah, N.H., Hastie, T., Tibshirani, R., . Some methods for heterogeneous treatment effect estimation in high dimensions. Statistics in Medicine 37, 1767–1787.
[65] Rahimi, A., Recht, B., . Random features for large-scale kernel machines, in: Advances in Neural Information Processing Systems.
[66] Rajkomar, A., Dean, J., Kohane, I., 2019. Machine learning in medicine. New England Journal of Medicine 380, 1347–1358.
[67] Robins, J., . 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, 1393–1512.
[68] Robins, J.M., Greenland, S., . The role of model selection in causal inference from non experimental data. American Journal of Epidemiology 123, 392–402.
[69] Robinson, P.M., . Root-n-consistent semiparametric regression. Econometrica , 931–954.
[70] Rolling, C.A., Yang, Y., . Model selection for estimating treatment effects 76, 749–769.
[71] Rosenbaum, P.R., Rubin, D.B., . The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.
[72] Rothman, K., Greenland, S., Lash, T., 2008. Case-control studies, chapter 8. Modern epidemiology , 111–127.
[73] Rubin, D.B., . Causal inference using potential outcomes. Journal of the American Statistical Association 100, 322–331.
[74] Saito, Y., Yasui, S., 2020. Counterfactual cross-validation: Stable model selection procedure for causal inference models, in: International Conference on Machine Learning, PMLR. pp. 8398–8407.
[75] Schulam, P., Saria, S., . Reliable decision support using counterfactual models. Advances in neural information processing systems 30.
[76] Schuler, A., Baiocchi, M., Tibshirani, R., Shah, N., . A comparison of methods for model selection when estimating individual treatment effects. arXiv:1804.05146 [cs, stat] .
[77] Schuler, M.S., Rose, S., . Targeted maximum likelihood estimation for causal inference in observational studies. American Journal of Epidemiology 185, 65–73.
[78] Shalit, U., Johansson, F.D., Sontag, D., 2017. Estimating individual treatment effect: generalization bounds and algorithms, in: International Conference on Machine Learning, PMLR. pp. 3076–3085.
[79] Shen, D., Wu, G., Suk, H.I., 2017. Deep learning in medical image analysis. Annual review of biomedical engineering 19, 221–248.
[80] Shen, L., Geleijnse, G., Kaptein, M., 2023. Rctrep: An r package for the validation of estimates of average treatment effects. Journal of Statistical Software .
[81] Shen, L., Margolies, L.R., Rothstein, J.H., Fluder, E., McBride, R., Sieh, W., 2019. Deep learning to improve breast cancer detection on screening mammography. Scientific reports 9, 12495.
[82] Shimoni, Y., Karavani, E., Ravid, S., Bak, P., Ng, T.H., Alford, S.H., Meade, D., Goldschmidt, Y., 2019. An evaluation toolkit to guide model selection and cohort definition in causal inference. arXiv preprint arXiv:1906.00442 .
[83] Shimoni, Y., Yanover, C., Karavani, E., Goldschmnidt, Y., . Benchmarking framework for performance-evaluation of causal inference analysis. arXiv:1802.05046 [cs, stat] .
[84] Simon, G.E., Johnson, E., Lawrence, J.M., Rossom, R.C., Ahmedani, B., Lynch, F.L., Beck, A., Waitzfelder, B., Ziebell, R., Penfold, R.B., et al., 2018. Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records. American Journal of Psychiatry 175, 951–960.
[85] Snowden, J.M., Rose, S., Mortimer, K.M., 2011. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. American Journal of Epidemiology 173, 731–738.
[86] Spasic, I., Nenadic, G., et al., 2020. Clinical text data in machine learning: systematic review. JMIR medical informatics 8, e17984.
[87] Sriperumbudur, B.K., Fukumizu, K., Gretton, A., Schölkopf, B., Lanckriet, G.R.G., . On integral probability metrics, \phi-

divergences and binary classification. arXiv:0901.2698 [cs, math] .

[88] Su, X., Peña, A.T., Liu, L., Levine, R.A., 2018. Random forests of interaction trees for estimating individualized treatment effects in randomized trials. Statistics in medicine 37, 2547–2560.

[89] Swaminathan, A., Joachims, T., 2015. Counterfactual risk minimization: Learning from logged bandit feedback, in: International Conference on Machine Learning, PMLR. pp. 814–823.

[90] VanderWeele, T.J., 2019. Principles of confounder selection. European journal of epidemiology 34, 211–219.

[91] Varoquaux, G., Colliot, O., 2022. Evaluating machine learning models and their diagnostic value.

[92] Wager, S., Athey, S., . Estimation and inference of heterogeneous treatment effects using random forests. Journal of the American Statistical Association 113, 1228–1242.

[93] Wang, H., Li, Y., Khan, S.A., Luo, Y., 2020. Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network. Artificial intelligence in medicine 110, 101977.

[94] Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N.H., Gallego, B., . Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases. Statistics in Medicine , 3309–3324.

[95] Yala, A., Lehman, C., Schuster, T., Portnoi, T., Barzilay, R., 2019. A deep learning mammography-based model for improved breast cancer risk prediction. Radiology 292, 60–66.

[96] Zadrozny, B., Elkan, C., . Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers , 8.

[97] Zhang, Z., Sejdić, E., 2019. Radiological images and machine learning: trends, perspectives, and prospects. Computers in biology and medicine 108, 354–370.

[98] Zivich, P.N., Breskin, A., 2021. Machine learning for causal inference: on the use of cross-fit estimators. Epidemiology (Cambridge, Mass.) 32, 393.

Appendix A. Variability of ATE estimation on ACIC 2016

Figure 1 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 [13].

Outcome models are implemented with scikit-learn [59] 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 A.2: Hyper-parameters grid used for ACIC 2016 ATE variability

Appendix B. Causal assumptions

We assume the following four assumptions, referred as strong ignorability and necessary to assure identifiability of the causal estimands with observational data [73]:

Assumption 1 (Unconfoundedness).

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

This condition –also called ignorability– is equivalent to the conditional independence on $e(X)$ [71]: $\{Y(0), Y(1)\} \perp\!\!\!\perp A|e(X)$.

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 [21], 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 [29].

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 [37].*

Appendix C. Proofs: Links between feasible and oracle risks

Appendix C.1. Upper bound of τ-risk with μ-risk$_{IPW}$

For the bound with the μ-risk$_{IPW}$, we will decompose the CATE risk on each factual population risks:

Definition 7 (Population Factual μ-risk). *[78]*

$$\mu\text{-}risk_a(f) = \int_{\mathcal{Y}\times\mathcal{X}} (y - f(x; A = a))^2 p(y; x = x \mid A = a)\, dydx$$

Applying Bayes rule, we can decompose the μ-risk on each intervention:

$$\mu\text{-risk}(f) = p_A\, \mu\text{-risk}_1(f) + (1 - p_A)\ \mu\text{-risk}_0(f) \text{with } p_A = \mathbb{P}(A = 1)$$

These definitions allows to state a intermediary result on each population:

Lemma 1 (Mean-variance decomposition). *We need a reweighted version of the classical mean-variance decomposition.*

For an outcome model $f : x \times A \to \mathcal{X}$. Let the inverse propensity weighting function $w(a; x) = ae(x)^{-1} + (1 - a)(1 - e(x))^{-1}$.

$$\int_{\mathcal{X}} (\mu_1(x) - f(x; 1))^2 p(x)dx = p_A \mu\text{-}risk_{IPW,1}(w, f) - \sigma^2_{Bayes}(1)$$

And

$$\int_{\mathcal{X}} (\mu_0(x) - f(x; 0))^2 p(x)dx = (1 - p_A)\mu\text{-}risk_{IPW,0}(w, f) - \sigma^2_{Bayes}(0)$$

Proof 1.

$$\begin{aligned}
&p_A \mu\text{-}risk_{IPW,1}(w, f) = \int_{\mathcal{X}\times\mathcal{Y}} \frac{1}{e(x)} (y - f(x; 1))^2 p(y \mid x; A = 1)p(x; A = 1)dydx \\
&= \int_{\mathcal{X}\times\mathcal{Y}} (y - f(x; 1))^2 p(y \mid x; A = 1) \frac{p(x; A = 1)}{p(x; A = 1)} p(x)dydx \\
&= \int_{\mathcal{X}\times\mathcal{Y}} \left[(y - \mu_1(x))^2 + (\mu_1(x) - f(x; 1))^2 + 2\,(y - \mu_1(x))\,(\mu_1(x) - f(x, 1))\,\right] p(y \mid x; A = 1)p(x)dydx \\
&= \int_{\mathcal{X}} \Big[\int_{\mathcal{Y}} (y - \mu_1(x))^2 p(y \mid x; A = 1)dy \Big] p(x)dx + \int_{\mathcal{X}\times\mathcal{Y}} (\mu_1(x) - f(x; 1))^2\, p(x)p(y \mid x; A = 1)dxdy \\
&+ \qquad 2 \int_{\mathcal{X}} \Big[\int_{\mathcal{Y}} (y - \mu_1(x))\, p(y \mid x; A = 1)dy \Big]\, (\mu_1(x) - f(x, 1))\, p(x)dx \\
&= \int_{\mathcal{X}} \sigma^2_y(x, 1)p(x)dx + \int_{\mathcal{X}} (\mu_1(x) - f(x; 1))^2\, p(x)dx + 0
\end{aligned}$$

Proposition 1 (Upper bound with mu-IPW). *Let f be a given outcome model, let the weighting function w be the Inverse Propensity Weight $w(x; a) = \frac{a}{e(x)} + \frac{1-a}{1-e(x)}$. Then, under overlap (assumption 2),*

$$\tau\text{-}risk(f) \leq 2\, \mu\text{-}risk_{IPW}(w, f)\ - 2\,(\sigma^2_{Bayes}(1) + \sigma^2_{Bayes}(0))$$

Proof 2.

$$\tau\text{-}risk(f) = \int_{\mathcal{X}} (\mu_1(x) - \mu_0(x) - (f(x; 1) - f(x; 0))^2 p(x)dx$$

By the triangle inequality $(u + v)^2 \leq 2(u^2 + v^2)$:

$$\tau\text{-}risk(f) \leq 2 \int_{\mathcal{X}} \left[(\mu_1(x) - f(x; 1))^2 + (\mu_0(x) - f(x; 0))^2 \right] p(x)dx$$

Applying Lemma 1,

$$\begin{aligned}
\tau\text{-}risk(f) &\leq 2\big[p_A \mu\text{-}risk_{IPW,1}(w, f) + \\
&(1 - p_A)\mu\text{-}risk_{IPW,0}(w, f)(w, f)\big] - 2(\sigma^2_{Bayes}(0) + \sigma^2_{Bayes}(1)) \\
&= 2\mu\text{-}risk_{IPW}(w, f) - 2(\sigma^2_{Bayes}(0) + \sigma^2_{Bayes}(1))
\end{aligned}$$

Appendix C.2. Reformulation of the R-risk as reweighted τ-risk

Proposition 2 (R-risk as reweighted τ-risk). ***Proof 3.*** *We consider the R-decomposition: [69],*

$$y(a) = m(x) + \big(a - e(x)\big)\tau(x) + \varepsilon(x;a) \tag{C.1}$$

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 3 : $\tilde{\sigma}^2_B(0) + \tilde{\sigma}^2_B(1)$

Appendix D. Measuring overlap

Motivation of the Normalized Total Variation. 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 [24] is typically used in the context of causal inference [78, 38]. 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 [87] 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 $\mathbb{P}(X|A)$:

$$\widehat{NTV}(e, 1-e) = \frac{1}{2N}\sum_{i=1}^{N} \Big|\frac{e(x_i)}{p_A} - \frac{1 - e(x_i)}{1 - p_A}\Big| \tag{D.1}$$

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

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

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

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

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 D.10.

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 D.12,
- Verifying that setups with penalized overlap from ACIC 2016 have a higher total variation distance than unpenalized setups in Figure D.11.
- 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 D.13. 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 D.10 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.

Appendix E. Experiments

Appendix E.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 [65]). We carefully control the overlap between treated and control populations, a crucial assumption for causal inference.

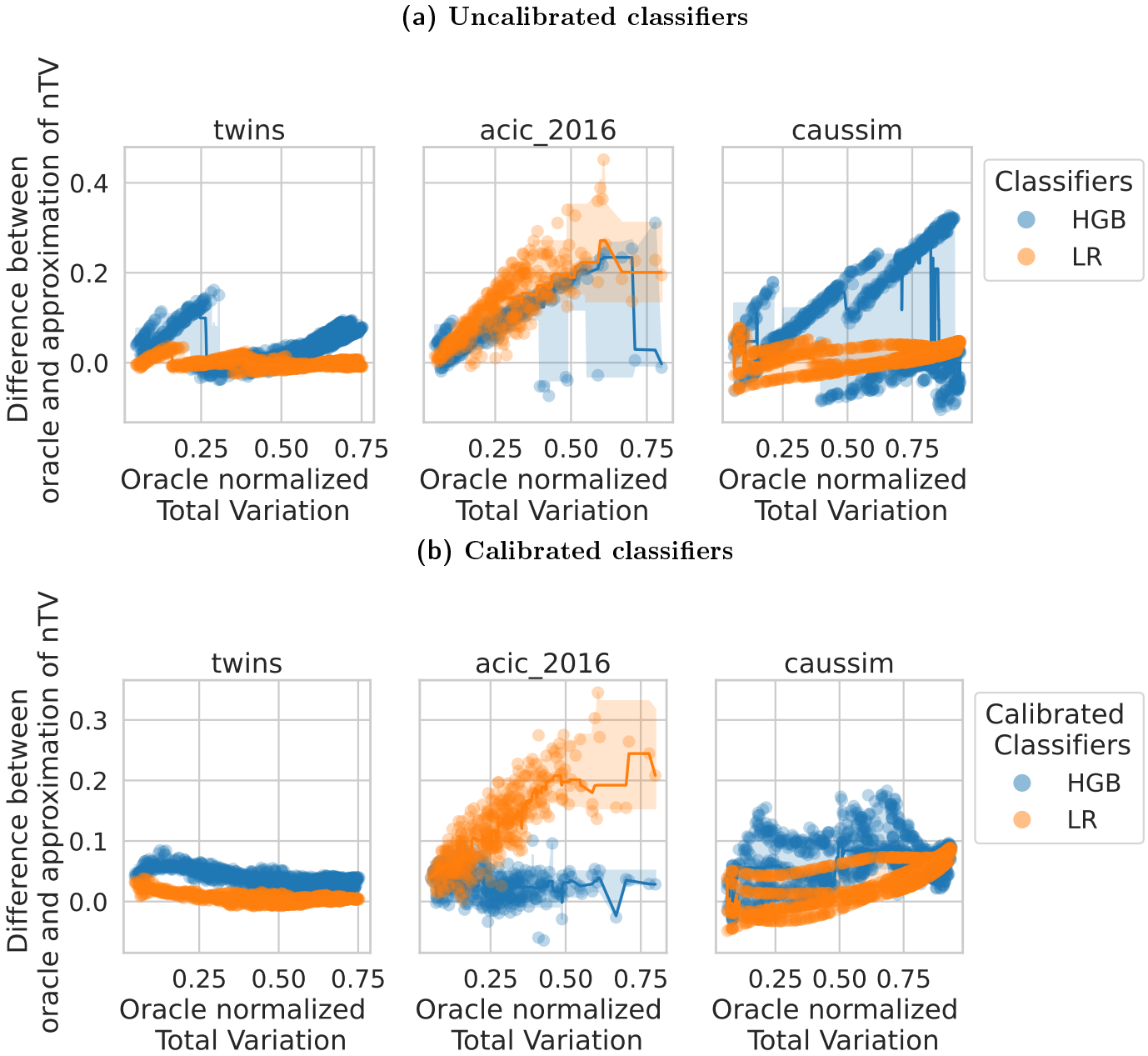


Figure D.10: a) Without calibration, estimation of NTV is not trivial even for boosting models. b) Calibrated classifiers are able to recover the true Normalized Total Variation for all datasets where it is available.

Figure D.11: NTV recovers well the overlap settings described in the ACIC paper [20]

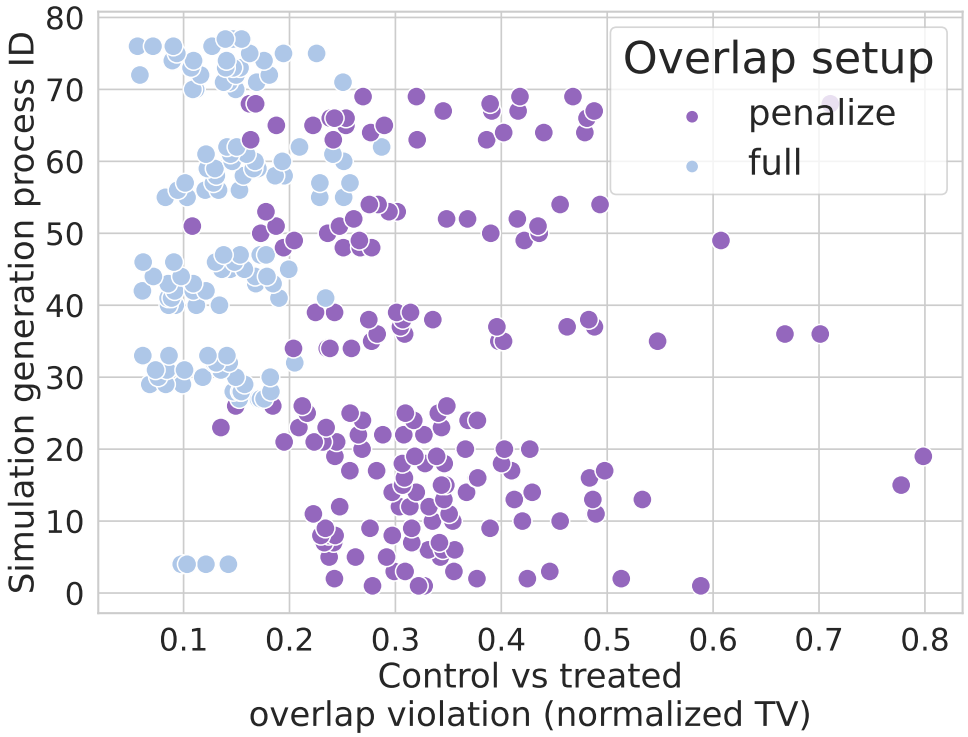
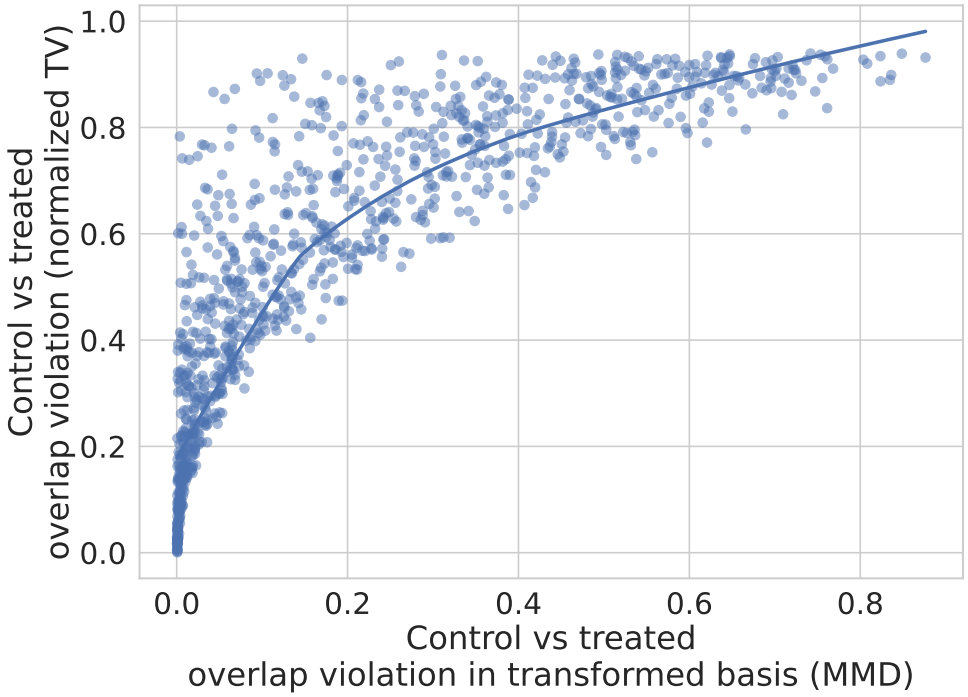


Figure D.12: Good correlation between overlap measured as normalized Total Variation and Maximum Mean Discrepancy (200 sampled Caussim datasets)



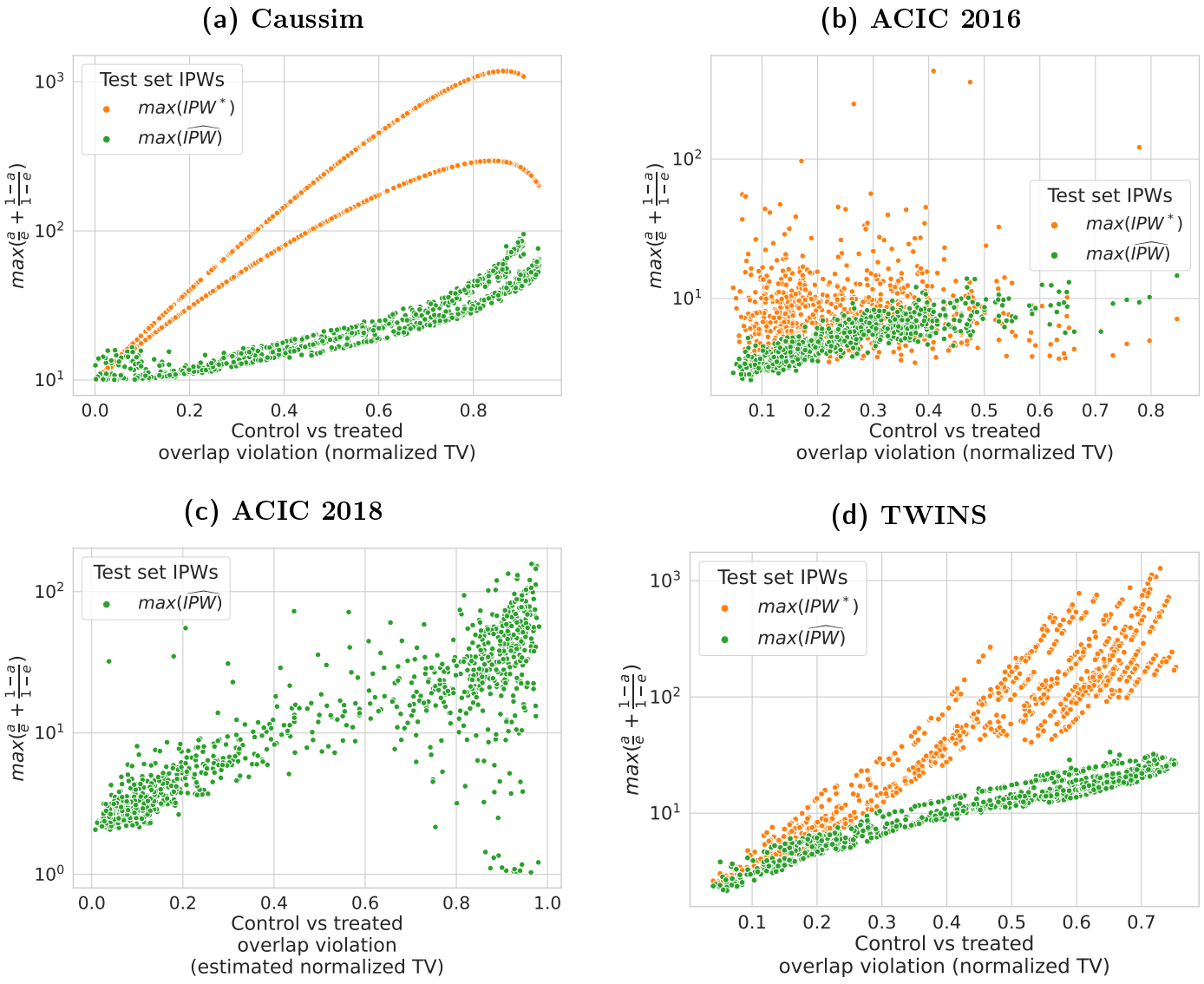


Figure D.13: Maximal value of Inverse Propensity Weights increases exponentially with the overlap as measure by Normalized Total Variation.

- 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:

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

  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 [8], 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 $z(x)$.

[8] We use the Sklearn implementation, [59]

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 $\theta \in [0, 2.5]$.

Appendix E.2. Model selection procedures

Nuisances estimation. The nuisances are estimated with a stacked regressor inspired by the Super Learner framework, [44]). The hyper-parameters are optimized with a random search with following search grid detailed in Table E.3. All implementations come from scikit-learn [59].

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, HistGradientBoostingRegressor max leaf nodes: [10, 20,
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, HistGradientBoostingClassifier max leaf nodes: [10, 20,

Table E.3: Hyper-parameters grid used for nuisance models

Appendix E.3. 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.

$$\kappa = \frac{(\text{number of concordant pairs }) - \ (\text{number of discordant pairs})}{(\text{number of pairs})} \tag{E.2}$$

Values of relative $\kappa(\ell, \tau-\text{risk})$ compared to the mean over all metrics Kendall's as shown in the boxplots of Figure 5.

Figure E.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 E.15 - Results measured in absolute Kendall's.

Figure E.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 E.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.

Metric	Dataset	Strong Overlap Median	Strong Overlap IQR	Weak Overlap Median	Weak Overlap 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 E.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 5

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

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

Selecting different seeds and parameters is crucial to draw conclucions. 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 E.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$_{IPW}$.

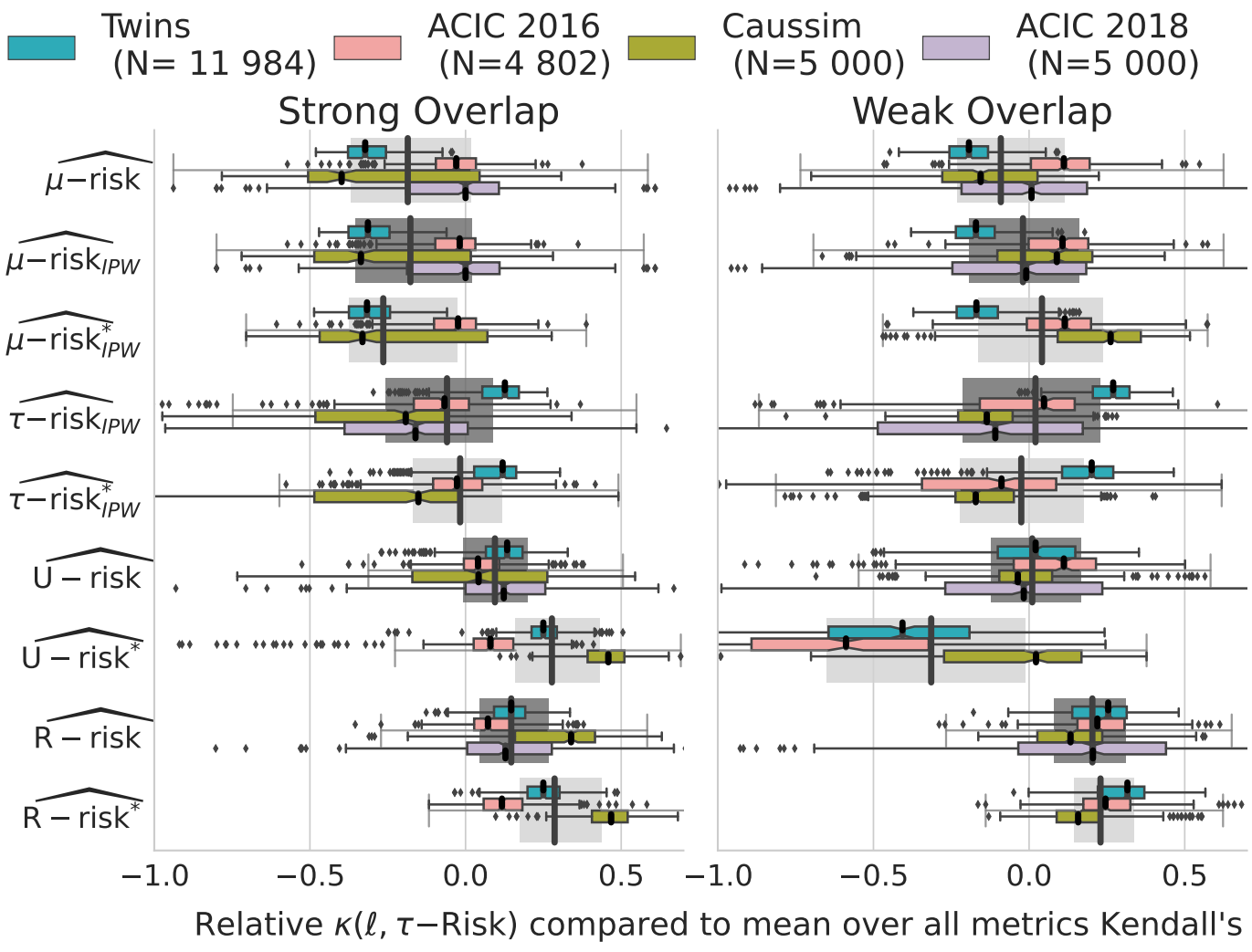


Figure E.14: **The R-risk is the best metric**: Relative Kendall's τ agreement with τ-risk. Strong and Weak overlap correspond to the first and last tertiles of the overlap distribution measured with Normalized Total Variation eq. D.1.

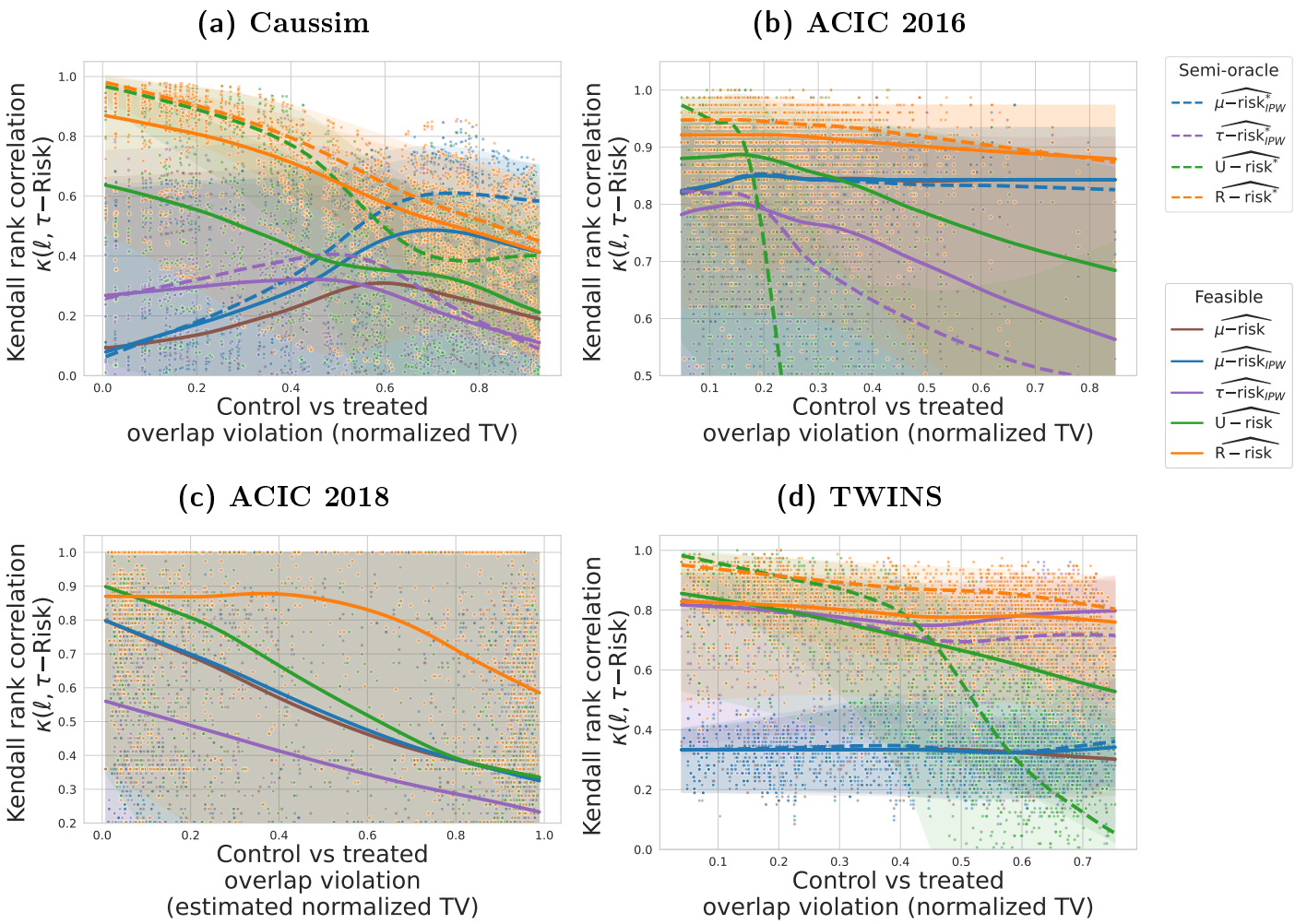


Figure E.15: Agreement with τ-risk ranking of methods function of overlap violation. The lines represent medians, estimated with a lowess. The transparent bands denote the 5% and 95% confidence intervals.

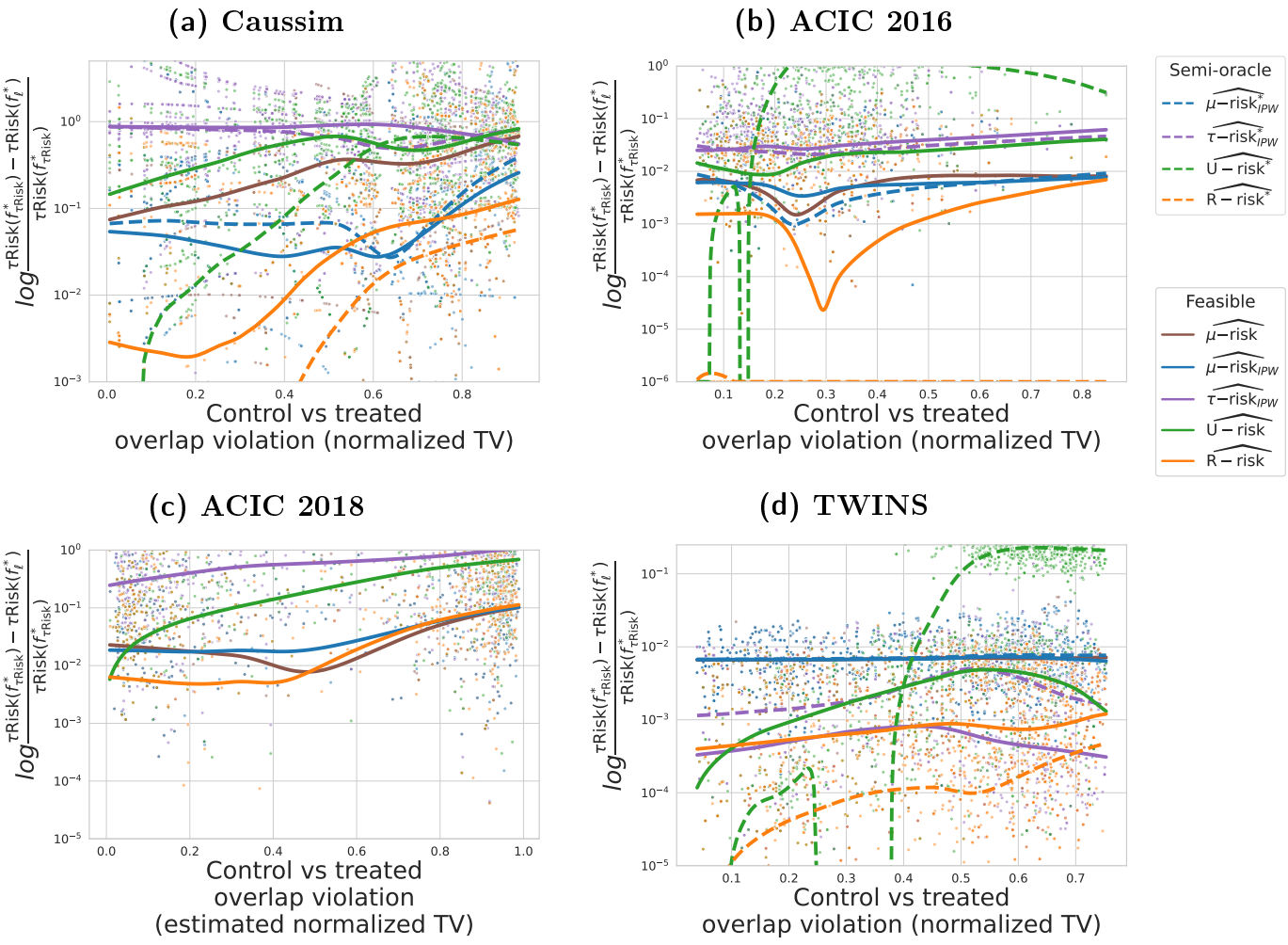


Figure E.16: Metric performances by normalized tau-risk distance to the best method selected with τ-risk. All nuisances are learned with the same estimator stacking gradient boosting and ridge regression. Doted and plain lines corresponds to 60% lowess quantile estimates. This choice of quantile allows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

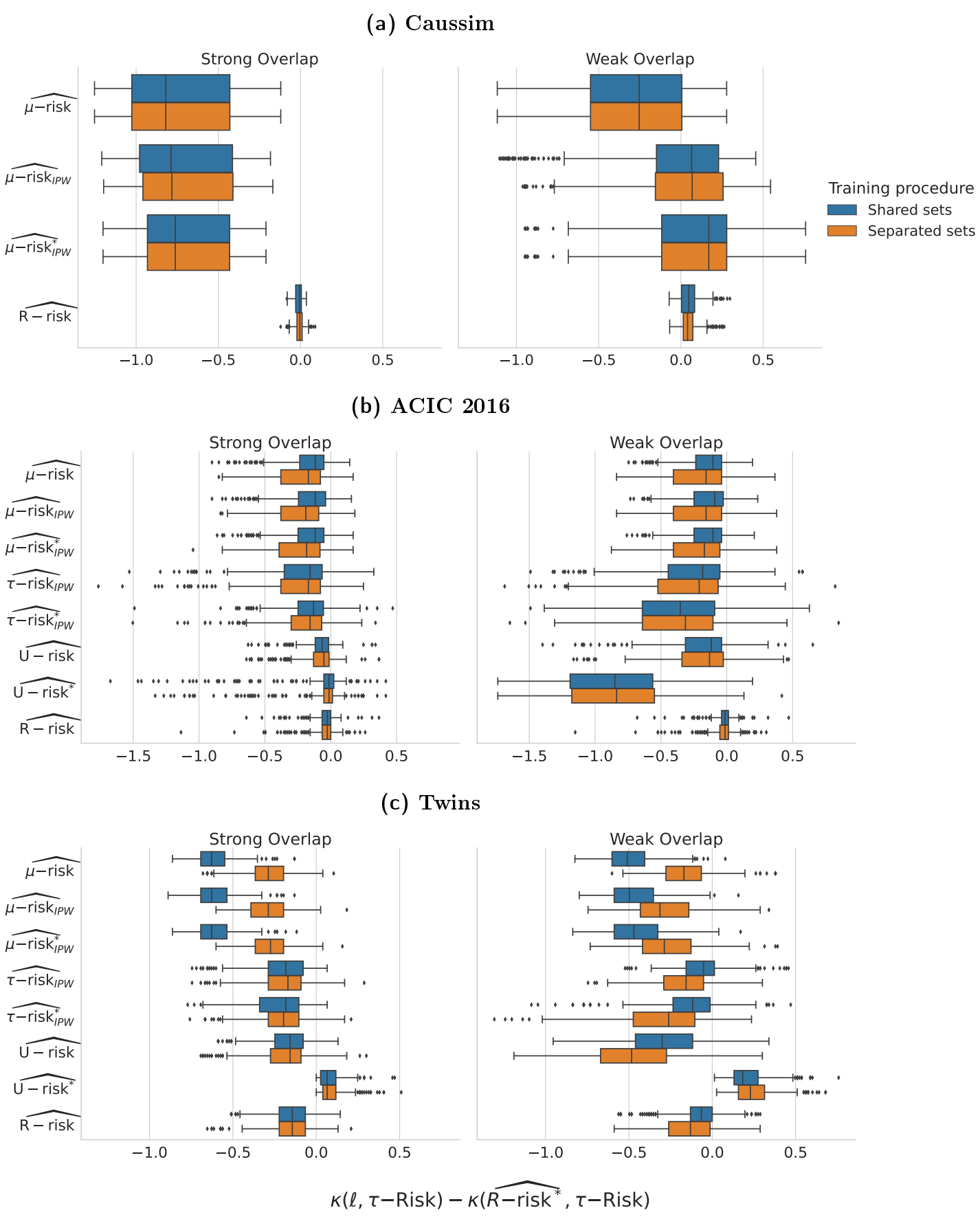


Figure E.17: Results are similar between the Shared nuisances/candidate set and the Separated nuisances set procedure. The experience has not been run on the full metrics for Caussim due to computation costs.

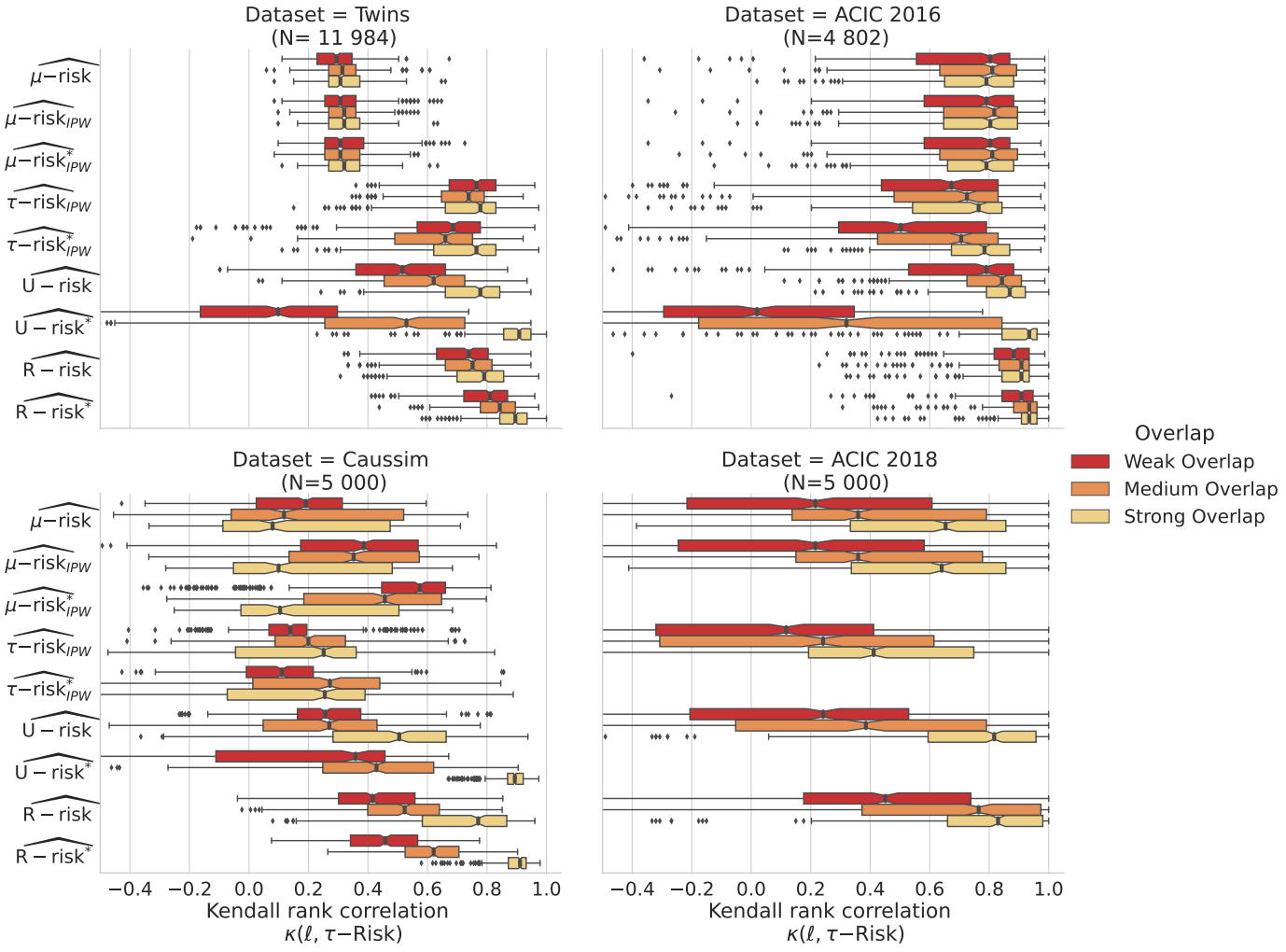


Figure E.18: **Low population overlap hinders causal model selection for all metrics**: Kendall's τ agreement with τ-risk. Strong, medium and Weak overlap correspond to the tertiles of the overlap distribution measured with Normalized Total Varation eq. D.1.

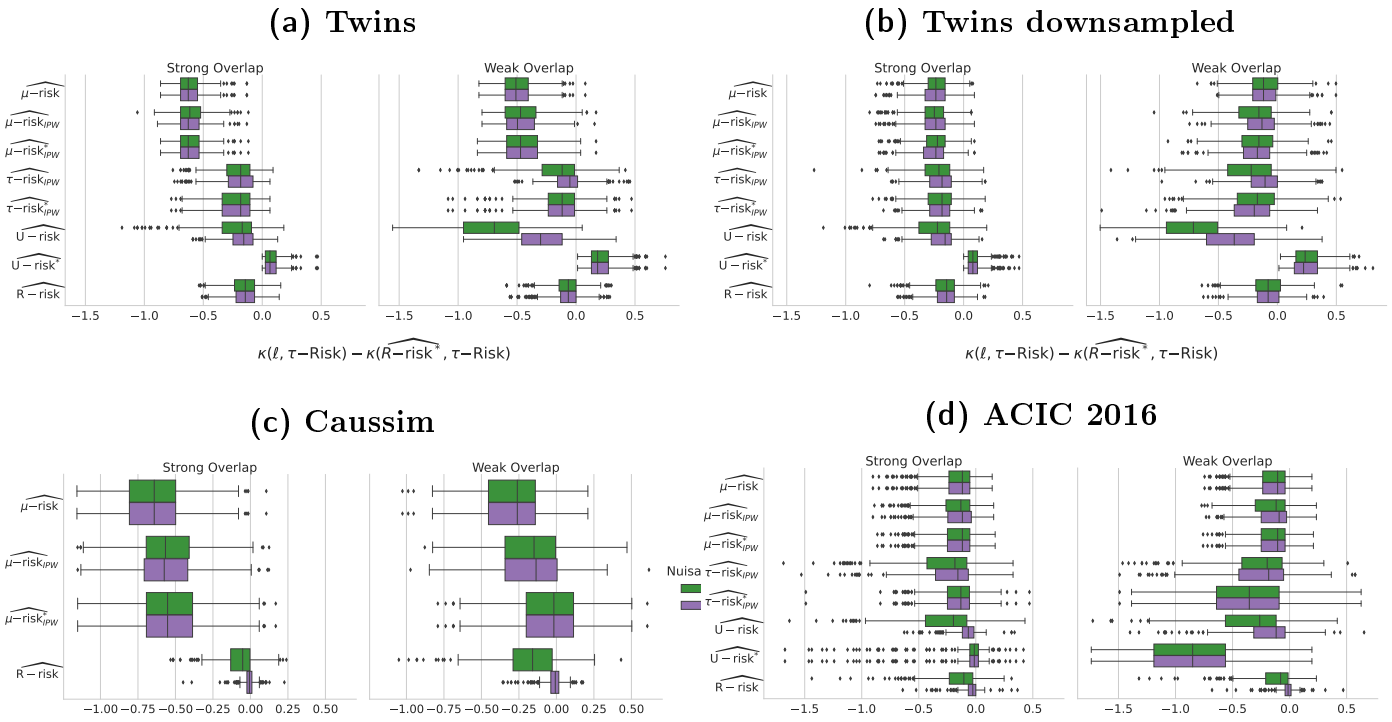


Figure E.19: Learning the nuisances with stacked models (linear and gradient boosting) is important for successful model selection with R-risk. For Twins dataset, there is no improvement for stacked models compared to linear models because of the linearity of the propensity model.

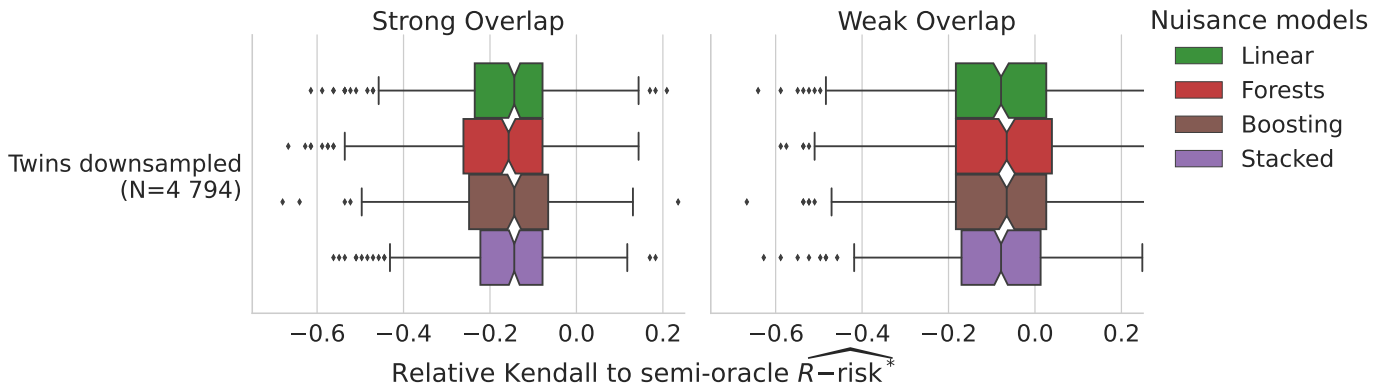
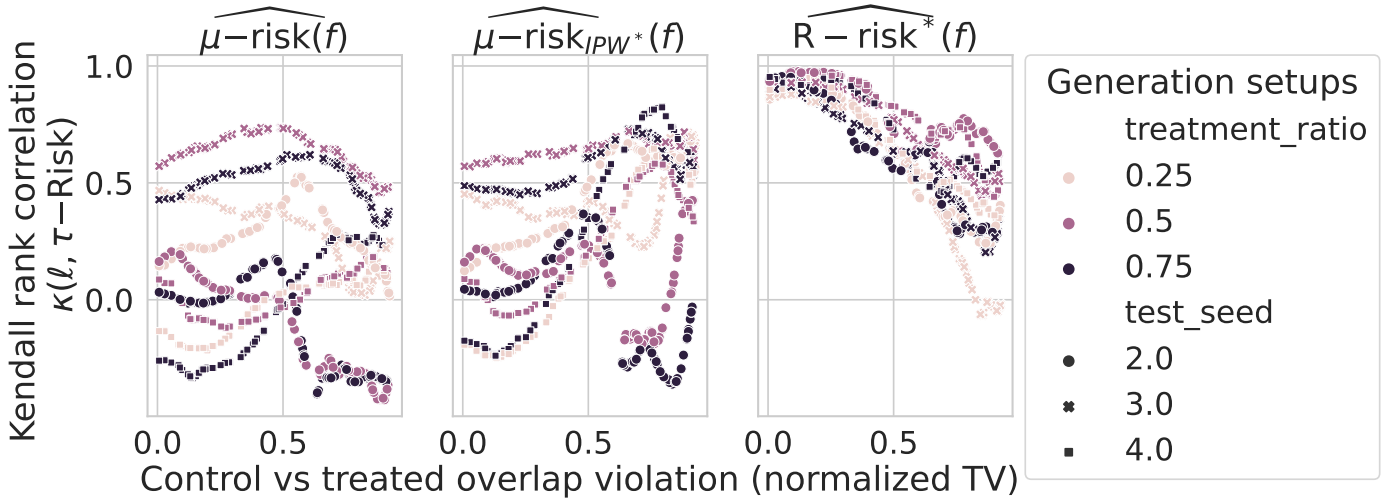


Figure E.20: **Flexible models are performant in recovering nuisances in the downsampled Twins dataset.** The propensity score is linear in this setup, making it particularly challenging for flexible models compared to linear methods.

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



Appendix F. 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 6). 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 [57]. Yet, it has been implemented with K=5 in causallib [82] or K=3 in econML [9]. Likewise, for ATE estimation, Chernozhukov et al. [15] 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. [47] use K=3, Gao et al. [23] use K=2. In the econML implementation, K is set to 3 [9]. Naimi et al. [53] 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 [26].

Causal model selection has been much less discussed. The only study that we are aware of, Schuler et al. [76], 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.