How to select predictive models for causal inference?

Matthieu Doutreligne, Gaël Varoquaux

How to select predictive models for causal inference?

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

[1]SoDa team, Inria, Saclay, France [2]Mission Data, Haute Autorité de Santé, Saint-Denis, France

*matthieu.doutreligne@inria.fr

Predictive models –as with machine learning– can underpin causal inference, to estimate the effects of an intervention at the population or individual level. This opens the door to a plethora of models, useful to match the increasing complexity of health data, but also the Pandora box of model selection: which of these models yield the most valid causal estimates? Classic machine-learning cross-validation procedures are not directly applicable. Indeed, an appropriate selection procedure for causal inference should equally weight both outcome errors for each individual, treated or not treated, whereas one outcome may be seldom observed for a sub-population. We study how more elaborate risks benefit causal model selection. We show theoretically that simple risks are brittle to weak overlap between treated and non-treated individuals as well as to heterogeneous errors between populations. Rather a more elaborate metric, the $R-$ risk appears as a proxy of the oracle error on causal estimates, observable at the cost of an overlap re-weighting. As the $R-$ risk is defined not only from model predictions but also by using the conditional mean outcome and the treatment probability, using it for model selection requires adapting cross validation. Extensive experiments show that the resulting procedure gives the best causal model selection.

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

1 | INTRODUCTION

1.1 | Valid causal inference from complex data requires causal model selection

There is growing interest in answering causal questions from observational data. While Randomized Control Trials (RCTs) remain the gold standard in medicine to estimate treatment effect[2], observational studies bring value to assess real-world effectiveness and safety[9], as they use the data from routine practice, or for drug repositioning[24,31] garnering first evidence without ethical concerns of systematic interventions[66]. The increasing amount of data collected routinely enables the use of increasingly flexible models that capture best heterogeneity and bridge to machine learning practices[45]. In particular the complexity of modern real-life health data, –Electronic Health Records, claims, or medical devices– calls for complex models.

For causal inference from observational data, epidemiology has historically focused on methods that model treatment assignment[33,56], based on the propensity score[4]. However, propensity-score methods are fragile to variance in probability estimates or lack of overlap between treated and non treated[21,65]. Recent empirical results[50,53] show a benefit of other types of methods, based on outcome modeling –also referred as G-computation or G-formula[6], Q-model in epidemiology[29] or conditional mean regression[50]. These outcome-modeling methods can easily go beyond Average Treatment Estimation (ATE), *eg* with Conditional Average Treatment Estimation (CATE), enabling to capture effect heterogeneity crucial for personalized medicine, to interpret the causal estimation on sub-populations, and policy optimization[62].

These methods capture the outcome as a function of the baseline covariates and the treatment with various models: Bayesian Additive Regression Trees[25], Targeted Maximum Likelihood Estimation[27,41], causal boosting[46], causal multivariate adaptive regression splines (MARS)[46], random forests[4952], Meta-learners[54], R-learners[39], Doubly robust estimation[43]... The wide variety of methods leaves the applied researcher with the difficult choice of selecting between different estimators based on the data at hand. Usual practices to select models in predictive settings rely on cross-validation on the error on the outcome[60,71]. In the case of causal inference, care must be taken that this error is not driven by inhomogeneities in treatment allocation. Indeed, while causal inference require modeling the links between an outcome and a treatment, the causal quantities are defined on a distribution distinct from the observed one: it includes *counterfactual* observations.

Given complex, potentially noisy, data, which model is to be most trusted to yield valid causal estimates? Because there is no single learner that performs best on all data sets, there is a pressing need for clear guidelines to select between causal models in health, economics and social science. Here we show that the best approach for model selection is to adapt cross-validation to estimate the so-called $R-$ risk which modulates observed prediction error to compensate for systematic differences between treated and non-treated individuals. The $R-$ risk relies on the two *nuisance* models, themselves estimated from data and thus imperfect; yet these imperfections do not undermine the benefit of the $R-$ risk.

1.2 | Prior work: model selection for outcome modeling (g-computation)

The natural risk for CATE model selection is a error measure between the true –unobserved– CATE (oracle) and the CATE estimate obtained with a candidate model of the outcome. But this risk is not "feasible": it cannot be computed solely from observed data and requires oracle knowledge.

Simulation studies of causal model selection

In simulations, the oracle CATE is known. Schuler et al. 2018[47] thus use eight simulation setups[46] to compare four causal risks, concluding that for CATE estimation the best model-selection risk is the R-risk[39] –def. 7, 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. 5– and reweighted MSE –μ-risk$_{IPW}$ def. 6– appear to perform better than R-risk on half of the simulations. Another work[51] studied empirically both MSE and reweighted MSE risks on the semi-synthetic ACIC 2016 datasets[53], but did not include the R-risk and looked only at the agreement of the best selected model with the true CATE risk –τ-risk(f) def. 4–, not on the full ranking of methods compared to the true CATE. Here we study experimentally a wider variety of data generative process for the observational setup. We also study the influence of overlap, an important parameter of the data generation process which makes a given causal metric appropriate[65].

Theoretical studies of causal model selection

Rolling and Yang 2014[32] propose a model selection procedure that asymptotically selects the best estimators among smooth models of the outcomes. However, practical cases often escape these theoretical requirement: it is delicate to assert whether

there are enough samples for asymptotic settings to hold –especially with high dimensionality– and candidate prediction models may not be smooth, as with popular tree-based methods.

Other work shows that unbiased estimates of the oracle CATE function $\tau(x)$ can be plugged into the oracle τ-risk for model selection. These CATE plugin estimators can be built with a simple IPW estimate[37], with a doubly robust estimator[61] or by debiasing a CATE estimator with influence functions[51] –in the like of Targeted Machine Learning[27,41]. However, theory holds for *well-specified* plugin CATE estimators and asymptotic regimes.

Statistical guarantees on causal estimation procedures

Much work in causal inference has focused on building procedures that guarantee asymptotically consistent estimators. Targeted Machine Learning Estimation (TMLE)[27,41] and Double Machine Learning[43] both provide estimators for Average Treatment Effect combining flexible treatment and outcome models. Here also, theories requires asymptotic regimes and at least assumes models to be *well-specified*.

By contrast, Johansson et al. 2021[67] 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 between the features of treated and control patients. However, they focus on using this upper bound for estimation, 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[47].

Objectives and structure of the paper

In this paper, we study *model selection procedures* (causal risks) in *finite samples* settings and without *well-specification* assumption. In these –practical– settings an important question is whether more complex risks, asymptotically consistent but with more quantities to estimate, suffer from more variance than their simpler though non-consistent counterparts, leading to worse model selection. In this respect, we compare semi-oracle settings, that use oracle knowledge of nuisance, to plugin estimates.

We first introduce the potential outcome framework and its notations, illustrating causal estimation with a toy example in Section 2. Then, we pose the causal model selection problem in Section 3, defining the studied causal risks. Section 4 gives our theoretical result. In section 5 we run a thorough empirical study, with many different settings covered. Finally, we comment our findings in Section 6.

2 | A CAUSAL-INFERENCE FRAMEWORK

2.1 | The Neyman-Rubin Potential Outcomes framework

Settings

Following the Neyman-Rubin Potential Outcomes framework[34], we observe 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 act, 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 $O^* = (Y(1), Y(0), X, A) \sim \mathcal{D}^* = \mathbb{P}(y(1), y(0), x, a)$.

At the population level, a popular quantity of interest –estimand– is the Average Treatment Effect (ATE), $\tau = \mathbb{E}_{Y(1),Y(0)\sim\mathcal{D}^*}[Y(1) - Y(0)]$. To model heterogeneity, the Conditional Average Treatment Effect (CATE), $\tau(x) = \mathbb{E}_{Y(1),Y(0)\sim\mathcal{D}^\star}[Y(1) - Y(0)|X = x]$, is also interesting.

Nuisances definitions

We define three important conditional expectancies required to estimate ATE and CATE but generally unknown. They are called nuisances in the causal inference literature, mostly in applied econometrics[43].

Definition 1 (Response surfaces)**.** The conditional expectancy of the outcome given the covariates and the treatment, $\mu_a(x) = \mathbb{E}_{Y\sim\mathcal{D}}[Y|X = x, A = a]$. It models the relation between the outcome and the patient characteristics in the observed distribution.

Definition 2 (Conditional mean outcome)**.** The conditional expectancy of the outcome given X, $m(x) = \mathbb{E}_{Y\sim\mathcal{D}}[Y|X = x]$. It marginalizes over the intervention, focusing on the link between the outcome and the covariates.

Definition 3 (Propensity score)**.** The conditional probability to be treated[4]: $e(x) = \mathbb{P}[A = 1|X = x]$. It models the intervention allocation.

Causal assumptions

Some assumptions are necessary to assure identifiability of the causal estimands in observational settings[16]. We assume the usual strong ignorability assumptions, composed of *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). In this work, we insist on the fundamental overlap assumption[65], which is testable with data.

Estimation 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 ignorability 1 assumptions imply the equality of two different expectations:

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

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[5]. For the rest of the paper, the expectations will always be taken on the factual observed distribution $\mathcal{D}$, and we will omit to explicitly specify the distribution. This identification leads to outcome based estimators (ie. g-computation estimators[29]), targeting the ATE τ with outcome modeling:

$$\tau = \mathbb{E}_{Y\sim D^\star}[Y(1) - Y(0)|X = x] = \mathbb{E}_{Y\sim D}[Y|A = 1] - \mathbb{E}_{Y\sim D}[Y|A = 0] \tag{2}$$

Given a sample of data and the oracle response functions μ_0, μ_1, the finite sum estimator of the ATE is written:

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

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, for the CATE, at the individual level:

$$\tau(x) = \mu_1(x) - \mu_0(x) \tag{4}$$

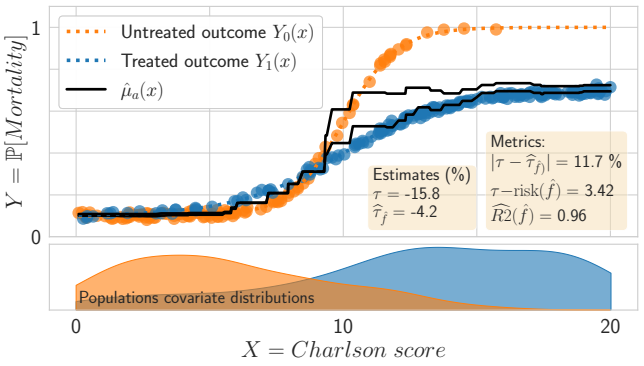
2.2 | Illustration: Toy example of causal model selection

Given various estimators of $\mu_0(x)$ and $\mu_1(x)$, we are interested in selecting those that minimize the estimation error on treatment effect. We illustrate that machine-learning model evaluation procedures such as Out-Of-Sample Mean Squared Error are not suited for this purpose. Figure 1 gives a toy example, with $Y \in [0, 1]$, the probability of death, a binary treatment $A \in \{0, 1\}$ and a single covariate $X \in \mathbb{R}$ which summarizes the patient health status (eg. the Charlson co-morbidity index[7]). We simulate a credible situation for which the treatment is beneficial (decreases the mortality probability) for patient with high Charlson scores (bad health states). On the contrary, the treatment has little effect for patients in good condition (small Charlson scores).

Some models of the response surfaces have high predictive performances of the outcome (measured as regression R2 score) but perform poorly for causal inference tasks such as Average Treatment Effect (error on the true effect τ) or Heterogeneous Treatment Effect inference (error on $\tau(x)$). Figure 1a shows a random forest with these counter-intuitive properties. On the contrary, Figure 1b shows a linear model with smaller R2 score but better causal inference.

Intuitively, the linear model misspecified –the outcome functions are not linear–, leading to poor R2; but it interpolates better to regions with poor overlap –high Charlson score– and thus gives better CATE estimates. Conversely, the random forest puts weaker assumptions on the data, thus has higher R2 score but is biased by the treated population in the poor overlap region, leading to bad CATE scores.

(a) Random forest model yielding bad inference



(b) Linear model yielding good inference

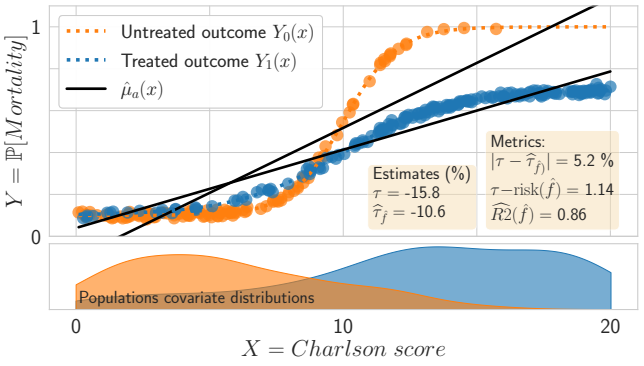


FIGURE 1 Toy example: (a) a random-forest estimator with high performance for standard prediction (high $\widehat{R2}$) but that yields poor ATE estimation (large error between true effect τ and estimated $\hat{\tau}_{\hat{f}}$), (b) a linear estimator with smaller prediction performance leading to better ATE and CATE estimation.

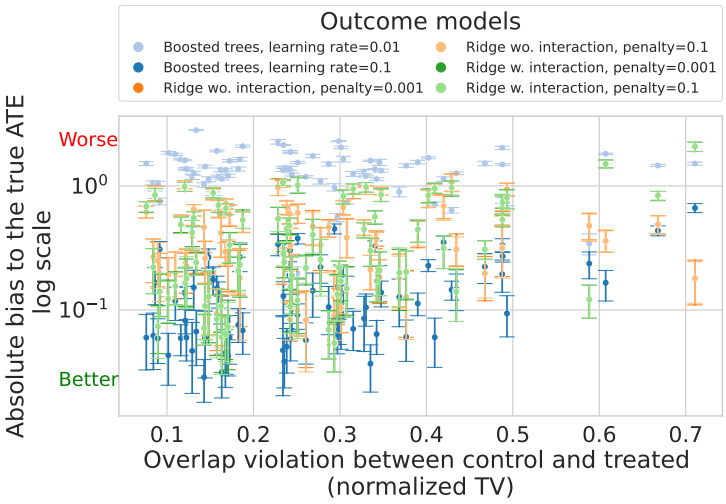
Selecting the estimator with the smallest τ-risk would lead to the smallest error on τ; however the τ-risk is not feasible: computing it requires access to unknown quantities.

While the random forest fits the data better than the linear model, it gives worse causal inference because its error is very inhomogeneous between the treated and untreated. The $\widehat{R2}$ score does not capture this inhomogeneity.

This toy example illustrates that the classic minimum Mean Square Error criterion is not suited to choosing a model among a family of candidate estimators for causal inference. Yet, model selection is a crucial aspect of causal inference. Indeed, estimates may vary markedly when using different models. For instance, figure 2 shows the large variations obtained across six different outcome estimators on the ACIC 2016 semi-synthetic datasets[53]. Flexible models such as boosting trees with a big learning rate (0.1) are doing well in most settings –in line with previous work[53]– except for setups with poor overlap, on the right of the plot. The same models with a small learning rate (0.01) yield the poorest performances. These two failure cases suggest that a simple rule of thumb such as preferring more flexible models does not work in general; an actual model-selection procedure is needed.

FIGURE 2 Average Treatment Effect estimations of six different outcome models used in g-estimators on the simulated data from the 76 simulations from ACIC 2016[53]. The models are boosted trees, ridge regression without interaction and ridge regression without interaction with the treatment. For each model, two choices of learning rate used during training are shown. The different configurations are plotted along with the overlap violation –measured with normalized Total Varation, def 15. Appendix A gives hyperparameter details.

We get non-consistent results with non overlapping error bars: choosing the best model among a family of candidate estimators is important.



3 | CAUSAL MODEL SELECTION: PROBLEM SETTING

3.1 | Causal model selection

We formalize the problem of 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 CATE and ATE:

$$\hat{\tau}_f(x) = f(x,1) - f(x,0) \quad \text{and} \quad \hat{\tau}_f(O) = \frac{1}{n}\sum_{i=1}^{n} \hat{\tau}_f(x_i) \tag{5}$$

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 metric of interest ℓ:

$$f^*_\ell = \underset{f \in \mathcal{F}}{\operatorname{argmin}}\, \ell(f, O) \tag{6}$$

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

3.2 | Model-selection risks, oracle and feasible

The τ-risk: an oracle error risk

Ideally, we would like to target the CATE, which naturally leads to the following evaluation risk:

Definition 4 (τ-risk(f))**.** also called PEHE[25,40]:

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

its finite-sum version over the observed data:

$$\widehat{\tau\text{-risk}}(f) = \sum_{x \in O} \left(\tau(x) - \hat{\tau}_f(x)\right)^2$$

However these risks are not feasible because the oracles $\tau(x)$ are not accessible, with the observed data $(Y, X, A) \sim \mathcal{D}$.

Feasible error risks

We explore **feasible risks**, based on the prediction error of the outcome model and *observable* quantities. Two of the following risks use the nuisances e –propensity score, def 3– and m –conditional mean outcome, def 2. 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 5 (Factual μ-risk(f))**.** [42] This is the usual Mean Squared Error on the target y. It is what is typically meant by "generalization error" in supervised learning and estimated with cross-validation:

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

Definition 6 (μ-risk$^\star_{IPW}(w, f)$)**.** [13] 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,

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

Definition 7 (R-risk$^\star(f)$)**.** [39,47] The R-risk uses the two nuisance m and e:

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

It has been introduced in causal-inference estimators motivated by its good approximation rate of τ, even with slow error rates on the nuisances $(\check{e}, \check{m})$[39].

These risks are summarized in Table 1.

TABLE 1 Review of causal risks

Risk	Equation	Reference
$mse(\tau(X), \tau_f(X)) = \tau\text{-risk}(f)$	$\mathbb{E}_{X\sim p(X)}[(\tau(X) - \hat{\tau}_f(X))^2]$	Eq. 4[25]
$mse(Y, f(X)) = \mu\text{-risk}(f)$	$\mathbb{E}_{(Y,X,A)\sim D}\left[(Y - f(X;A))^2\right]$	Def. 5[47]
$\mu\text{-risk}^*_{IPW}$	$\mathbb{E}_{(Y,X,A)\sim D}\left[\left(\frac{A}{e(X)} + \frac{1-A}{1-e(X)}\right)(Y - f(X;A))^2\right]$	Def. 6[13]
$R\text{-risk}^*$ [1]	$\mathbb{E}_{(Y,X,A)\sim D}\left[\left((Y - m(X)) - (A - e(X))\,\tau_f(X)\right)^2\right]$	Def. 7[39]

[1] Called $\tau\text{-risk}_R$ in Schuler et al. 2018[47].

3.3 | Estimation and model selection procedure

Causal model selection (as in *eg* Equation 6) may involve estimating a variety of 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 $(\mathcal{T}, \mathcal{S})$ of sizes $\left(\frac{N}{2}, \frac{N}{2}\right)$. 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 illustrated in Figure 3.

As extreme inverse propensity weights induce high variance, clipping can be usefull to ensure numerical stability[18,35].

Using the train set $\mathcal{T}$ both to fit the candidate estimator and the nuisance estimates is a form of double dipping which leads to correlation in the final estimates[39]. However, comparing with a procedure where the nuisances are learned on a separated validation set, did not reveal important changes to the final results (see appendix E.2). We thus kept this simple two-sets procedure.

Algorithm 1 Evaluation of selection procedures for one simulation

Given a train and a test sets $(\mathcal{T}, \mathcal{S}) \sim \mathcal{D}$, a family of candidate estimators $\{f \in \mathcal{F}\}$, a set of causal metrics $\ell \in \mathcal{L}$:

1. Prefit: Learn estimators for unknown nuisance quantities $(\check{e}, \check{m})$ on the training set $\mathcal{T}$
2. Fit: $\forall f \in \mathcal{F}$ 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 each candidate estimator with each causal metric $\mathcal{M}(\hat{f}, \mathcal{S})$. For each causal metric $\ell \in \mathcal{L}$ and each candidate estimator $f \in \mathcal{F}$, store the metric value: $\ell(f, \mathcal{S})$ – possibly function of $\check{e}$ and $\check{m}$

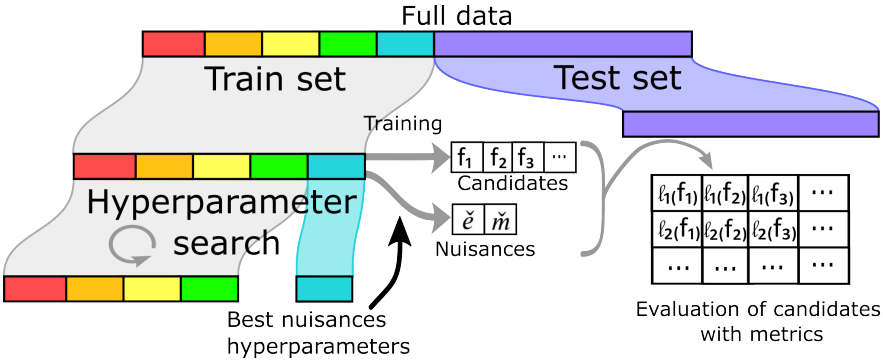


FIGURE 3 Estimation procedure for causal model selection.

4 | THEORY: LINKS BETWEEN FEASIBLE AND ORACLE RISKS

We recall that the μ-risk$_{IPW}$ can upper bound the oracle τ-risk. We show that the R-risk appears as a reweighted version of the oracle τ-risk. Both results make explicit the role of overlap for the performances of causal risks.

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[42]:

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

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.

4.1 | Upper bound of τ-risk with μ-risk$_{IPW}$

Proposition 1 (Upper bound with μ-risk$_{IPW}$)**.** [67] 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 already been derived in previous work[67]. It links μ-risk$_{IPW}$ to the squared residuals of each population thanks to a reweighted mean-variance decomposition. For completeness, we provide the proof in Appendix C.1.

The upper-bound comes from the triangular inequality applied to the residuals of both populations. Interestingly, the two quantities are equal when the absolute residuals on treated and untreated populations are equal on the whole covariate space, *ie* for all $x \in \mathcal{X}$, $|\mu_1(x) - f(x, 1)| = |\mu_0(x) - f(x, 0)|$. The main source of difference between the oracle τ-risk and the reweighted mean squared error, μ-risk$_{IPW}$, comes from heterogeneous residuals between populations. These quantities are difficult to characterize as they are linked both to the estimator and to the data distribution. This bound indicates that minimizing the μ-risk$_{IPW}$ helps to minimize the τ-risk, which leads to interesting optimization procedures[67]. However, there is no guarantee that this bound is tight, which makes it less useful 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: $\tau\text{-risk}(f) \leq \frac{2}{\eta} \, \mu\text{-risk}(f) - 2 \left(\sigma_B^2(1) + \sigma_B^2(0)\right)$. For weak overlap (propensity scores not bounded far from 0 or 1), this bound is very loose (as shown in Figure 1) and is not appropriate to discriminate between models with close performances.

4.2 | 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):

$$R\text{-risk}^*(f) = \int_x e(x)\big(1 - e(x)\big)\big(\tau(x) - \tau_f(x)\big)^2 p(x) dx \; + \; \tilde{\sigma}_B^2(1) \; + \; \tilde{\sigma}_B^2(0) \tag{7}$$

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 goes down to zero with the propensity score.

4.3 | 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 $\mu\text{-risk}_{IPW}$ 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, even if we have randomization, we still can have large differences between τ-risk and $\mu\text{-risk}_{IPW}$ coming from heterogeneous errors between populations as noted in Section 4.1 and shown experimentally in simulations[47].

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:

$$R\text{-risk} = p_A\,(1-p_A)\,\tau\text{-risk} \;+\; (1-p_A)\,\sigma_B^2(0) \;+\; p_A\sigma_B^2(1) \tag{8}$$

Therefore, optimizing estimators for CATE with R-risk* in the randomized setting is optimal if we target the τ-risk. This explains the strong performances of R-risk in randomized setups[47] and is a strong argument in favor of this risk for heterogeneity estimation in RCTs.

Oracle Bayes predictor

Consider the case where 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:

$$\tau\text{-risk}(\mu) = \mathbb{E}_{X\sim p(X)}[(\tau(X) - \tau_\mu(X))^2] = 0 \tag{9}$$

$$\mu\text{-risk}(\mu) = \mathbb{E}_{(Y,X,A)\sim p(Y;X;A)}[\big(Y - \mu_A(X)\big)^2] = \int_{\mathcal{X},\mathcal{A}} \varepsilon(x,a)^2 p(a \mid x)\, p(x)\, dx\, da \leq \sigma_B^2(0) + \sigma_B^2(1) \tag{10}$$

$$\mu\text{-risk}_{IPW}(\mu) = \sigma_B^2(0) + \sigma_B^2(1) \quad \text{follows from Lemma 1} \tag{11}$$

$$R\text{-risk}(\mu) = \tilde{\sigma}_B^2(0) + \tilde{\sigma}_B^2(1) \leq \sigma_B^2(0) + \sigma_B^2(1) \quad \text{follows directly from Proposition 2} \tag{12}$$

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. However, in practice choices must be made in non-asymptotic regimes.

5 | EMPIRICAL STUDY

We evaluate the following causal metrics, oracle and feasible versions of finite-sample evaluation risks presented in Table 1:

$$\mathcal{L} = \left\{ \widehat{\mu\text{-risk}}^{*}_{IPW},\ \widehat{R\text{-risk}}^{*},\ \widehat{\mu\text{-risk}},\ \widehat{\mu\text{-risk}}_{IPW},\ \widehat{R\text{-risk}} \right\} \tag{13}$$

We compare them on a large sample of different simulated data generation processes to select best performing estimator among a family of candidate estimators. We also evaluate them on three semi-simulated datasets: ACIC 2016[53], ACIC 2018[48] and Twins[38]. [1]

5.1 | Extensive simulation settings

Data Generation Process

We use simulated data, on which the ground-truth causal effect is known. Going further than prior empirical studies of causal model selection[47,51], we use multiple generative processes, to reach conclusions wider than a given one (as discussed in Appendix E10).

[1] Scripts for the simulations and the selection procedure are available at https://github.com/soda-inria/caussim. Results of the main experience described in this section are also provided to avoid re-running the full experience.

We generate random functions for the response functions using random bases. Basis extension methods are common in biostatistics where spline are often used for functional regression[26,55]. By allowing the function to vary at specific knots, they give flexible –non-linear– models of the studied mechanisms. Taking inspiration from splines, we use random approximation of Radial Basis Function (RBF) kernels[19] to generate the response surfaces. RBF use the same process as polynomial splines but replace polynomial by Gaussian kernels. Unlike polynomials, Gaussian kernels have exponentially decreasing influences in the input space. This allows to avoid unrealistic divergences of the population 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 amount of overlap between treated and control populations, as it an important parameter for causal inference (related to η which appears in section 4.1). Finally, we generate the observed outcomes adding some Gaussian noise. We generated such datasets 1000 times, with uniformly random overlap parameters $\theta \in [0, 2.5]$. Appendix E.1 gives more details on the data generation.

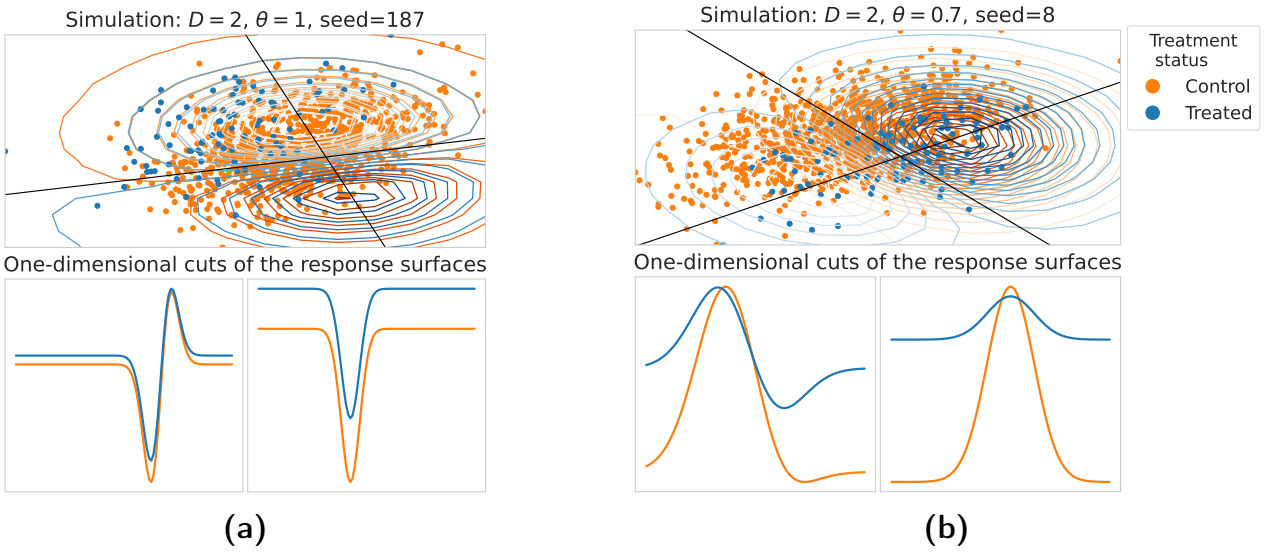


FIGURE 4 Two examples of the simulation setup in the input space with two knots –*ie.*basis functions: with low 4a and high 4b overlap setups. The top row gives views of the observations in feature space, while the lower row displays the two response surfaces on a 1D cut along the black lines drawn on the above panel.

Family of candidate estimators

We build a candidate estimator in two steps. First, we use a RBF expansion similar as the one used for the data-generation generation process. Concretely, we choose two random knots and apply a transformation of the raw data features with the same Gaussian kernel used for the data-generation mechanism. This step is referred as the featurization. Then, we fit a linear regression on this transformed features. We consider two ways of combining these steps for outcome mode; using common nomenclature[54], we refer to these regression structures as different meta-learners which 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 representation.
- TLearner: Two completely different learners for each population, hence separate featurization and separate regressors.

We are not including more elaborated meta-learners such as R-learner[39] or X-learner[54]. Our goal is not to have the best possible learner but to have a variety of sub-optimal learners in order 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 the learning procedure and the featurization yielding a family $\mathcal{F}$ of 120 candidate estimators.

5.2 | Semi-simulated datasets

Datasets

We also use semi-simulated datasets, where a known synthetic causal effect is added to real –non synthetic– covariate. We use datasets used in previous work to evaluate causal inference:

- ACIC 2016[53]: The dataset is based on the Collaborative Perinatal Project[3], a RCT conducted on a cohort of pregnant women to identify causes of 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 given 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 with varying parameters for treatment and response generation models, treatment assignment probabilities, overlap, and interactions between treatment and covariates [2]. We used 10 different seeds for every setup, totalizing 770 dataset instances.
- ACIC 2018[48]: The raw covariates data comes from the Linked Births and Infant Deaths Database (LBIDD)[10] with $D = 177$ covariates. Treatment and outcome models has been simulated with complex models in order to reflect different scenarii of inference. They do not provide the true propensity scores, so we evaluate only the feasible metrics which does not require this nuisance parameter. We used all datasets of size $N = 5\,000$, totalizing 432 dataset instances [3].
- Twins[38]: It is an augmentation of the real data on twin births and mortality rates in the USA from 1989-1991[14]. There are $N = 11\,984$ samples (pairs of twins), and $D = 50$ covariates, The outcome is the mortality and the treatment is the weight of the heavier twin at birth. This is a "true" counterfactual dataset –as remarked in[64]– in the sense that we have both potential outcomes with each twin. They simulate the treatment with a sigmoid model based on GESTAT10 (number of gestation weeks before birth) and x the 45 other covariates:

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

  We built upon this equation, adding a non-constant slope in the treatment sigmoid, allowing us to control the amount of overlap between treated and control populations. [4] We sampled uniformly 1 000 different overlap parameters between 0 and 2.5, totalizing 1 000 dataset instances. Unlike the previous datasets, only the overlap varies for these instances. The response surfaces are fixed by the original twin 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[41], 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 (detailed of the hyper parameters in Appendix E.2). As extreme inverse propensity weights induce high variance, we use clipping[18,35] to bound $min(\check{e}, 1 - \check{e})$ away from 0 with a fixed $\eta = 10^{-10}$, ensuring strict overlap for numerical stability.

5.3 | Measuring overlap between treated and non treated

Overlap between treated and control population is crucial for causal inference, it appears in the positivity assumption 2 required for causal identification and when relating the different risks (subsection 4.1).

[2] Original R code available at https://github.com/vdorie/aciccomp/tree/master/2016 to generate 77 simulations settings.
[3] Using only the scaling part of the data, obtained from the https://github.com/IBM-HRL-MLHLS/IBM-Causal-Inference-Benchmarking-Framework
[4] We obtained the dataset from https://github.com/AMLab-Amsterdam/CEVAE/tree/master/datasets/TWINS
[5] Scikit-learn regressor, HistGradientBoostingRegressor, and classifier, HistGradientBoostingClassifier.

Overlap, or "positivity", is typically assessed by qualitative methods using population histograms (as in Figure 1) or side-by-side box plots, or quantitative approaches such as Standardized Mean Difference[23,33]. While these methods are useful to decide if positivity holds they do not summarize a dataset's overlap in a single measure. Rather, divergence between distributions $\mathbb{P}(X|A=0)$ and $\mathbb{P}(X|A=1)$ give a relevant quantity to characterize the behavior of causal risk[65,67].

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 [6]. This is the empirical measure of the total variation distance[22] between the distributions, $TV(\mathbb{P}(X|A=1), \mathbb{P}(X|A=0))$. As we have both distribution sampled on the same points, we can rewrite it a sole function of the propensity score, a low dimensional score more tractable than the full distribution $\mathbb{P}(X|A)$:

$$\widehat{NTV}(e, 1-e) = \frac{1}{2N}\sum_{i=1}^{N}\left|\frac{e(x_i)}{p_A} - \frac{1-e(x_i)}{1-p_A}\right| \tag{15}$$

Appendix D gives a detailed theoretical motivation of the NTV distance and empirical arguments showing that it recovers the desired notion of overlap.

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 D1.

5.4 | Empirical results

We investigate how well the various causal metrics rank the different candidate models. Figure 5 shows the Kendall rank correlation coefficient[1] between the ranking of methods given the oracle τ-risk and every causal metric under evaluation. We plot this percentage of agreement as a function of decreasing overlap (by increasing Normalized Total Variation).

R-risk dominates factual μ-risk and its reweighted version

Among all causal metrics, classical mean squared error (ie. factual μ-risk) is suboptimal. Reweighting it with propensity score (μ-risk$_{IPW}$) does not bring much improvements. Including a model of the outcome in the *R*-risk leads to better performances in every cases. Further results provided in Appendix E.3 with alternative measures of performance confirm these findings.

Low population overlap hinders causal model selection

The performances of every metric drop with growing lack of overlap. It is particularly visible for Caussim, ACIC 2018 and Twins. Model selection for causal inference becomes more and more difficult with increasingly different treated and control populations.

Estimating the nuisances does not hinder model selection

Oracle versions of every risks recover more often the best estimator. However, flexible nuisances estimations (gradient boosting trees) lead to feasible metrics with close performances to the oracles ones. This suggests that the chosen estimators are doing well in recovering the true nuisances.

[6] Computing overlap when working only on samples of the observed distribution, outside of simulation, requires a more sophisticated estimator of discrepancy between distributions, as two data points never have the same exact set of features. Maximum Mean Discrepancy[30] is typically used in the context of causal inference[42,67]. 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.

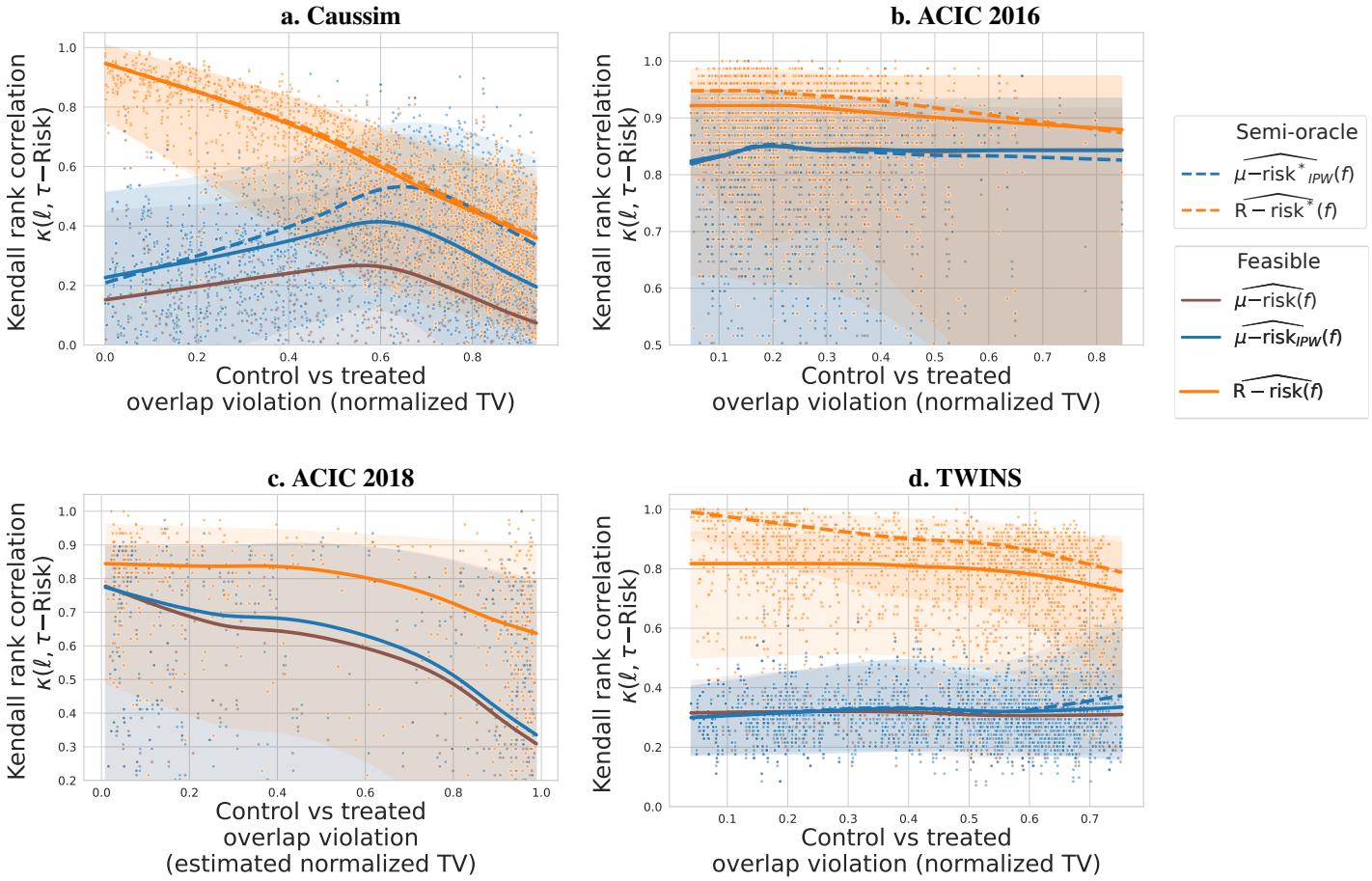


FIGURE 5 Agreement with τ-risk ranking of methods function of overlap violation. The lines represent medians, estimated with a lowess. The transparent bands denote the 5% and 95% confidence intervals.

6 | DISCUSSION AND CONCLUSION

Predictive models are increasingly used to reason about causal effects. 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. 7) leads to more valid causal estimates. Estimating this risk requires a markedly more complex procedure than standard cross-validation used *e.g.* in machine learning: it involves fitting nuisance models necessary for model evaluation, though our empirical results show that these can be learned on the same set of data as the outcome model evaluated. A poor estimation of the nuisance models may compromise the benefits of the more complex R − risk (as shown in in Appendix E9). However controlling and selecting these latter models is easier because they are associated to errors on observed distributions and our empirical results show that when selecting these models in a flexible family of models the R − risk dominates simpler risks for model selection. Our results show that going from an oracle R − risk –where the nuisances are known– to a feasible R − risk –where the nuisances are estimated– decreases only very slightly the model-selection performance of the R − risk. This may be explained by theoretical results that suggest that estimation errors on both nuisances partly compensate out in the R − risk[39,43,44,59,69]. The usage of the R-risk can also be understood as a τ-risk reweighted by the propensity score (prop 2).

For strong overlap, the μ − risk appears theoretically motivated (subsection 4.1), however empirical results show that even in this regime the R − risk brings a sizeable benefit, in agreement with Schuler et al. 2018[47].

Extension to binary outcome

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[36]; in particular the odds ratio can carry across different disease sampling rates[20]. Using as an estimand the log of these values is suitable to additive models (for reasoning or noise assumptions). In the log domain, the relative risk or the odds ratio are written as a difference, as the ATE: $\log \mathbb{P}(Y(1)=1)-\log \mathbb{P}(Y(0=1)$ or $\log(\mathbb{P}(Y(1)=1)/[1-\mathbb{P}(Y(1)=1)])-\log \mathbb{P}(Y(0=1)/[1-\mathbb{P}(Y(0)=1)]$. Hence, the framework studied here (subsection 2.1) can directly apply. It is particularly easy for the log odds ratio, as it is the output of a logistic regression or any model with a cross-entropy loss.

Going further

The $R-$ risk needs good estimation of nuisance models. The propensity score e calls for a control on the estimation of the individual posterior probability. We have used the Brier score to select these models, as it is minimized by the true individual probability. Regarding model-selection for propensity score, an easy mistake is to use expected calibration errors popular in machine learning[11,12,15,68] as these select not for the individual posterior probability but for an aggregate error rate[70]. An open question is whether a better metric than the brier score can be designed that controls for $e\,(1-e)$, the quantity used in the $R-$risk, rather than e.

The quality of model selection varies substantially from one data-generating mechanism to another. The overlap appears as an important parameter: when the treated and untreated, causal model selection is very hard. However, remaining variance in the empirical results suggests that other parameters of the data generation processes come into play. Intuitively, the complexity of the response surfaces and the treatment heterogeneity interact with overlap violations: when extrapolations to weak-overlap regions is hard, causal model selection is hard.

Nevertheless, from a practical perspective, our study establishes that the R-risk is the best option to select predictive models for causal inference, without requiring assumptions on the data-generating mechanism, the amount of data at hand, or the specific estimators used to build predictive models.

ACKNOWLEDGMENTS

We acknowledge fruitful discussions with Bénédicte Colnet.

References

(1) Kendall, M. G. A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93.

(2) OR, Jones; WD,Platt Streptomycin Treatment of Pulmonary Tuberculosis. *British Medical Journal* **1948**, *2*, 769–782.

(3) 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*, Google-Books-ID: A0bdVhlhDQkC; National Institute of Health: 1972; 562 pp.

(4) Rosenbaum, P. R.; Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika* **1983**, *70*, 41–55.

(5) 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* **1986**, *7*, 1393–1512.

(6) Robins, J. M.; Greenland, S. The role of model selection in causal inference from non experimental data. *American Journal of Epidemiology* **1986**, *123*, 392–402.

(7) 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* **1987**, *40*, 373–383.

(8) Robinson, P. M. Root-N-Consistent Semiparametric Regression. *Econometrica* **1988**, *56*, Publisher: [Wiley, Econometric Society], 931–954.

(9) Black, N. Why we need observational studies to evaluate the effectiveness of health care. *Bmj* **1996**, *312*, 1215–1218.

(10) MacDorman, M. F.; Atkinson, J. O. Infant mortality statistics from the linked birth/infant death data set–1995 period data. *Monthly Vital Statistics Report* **1998**, *46*, 1–22.

(11) Platt, J. C.; Platt, J. C. Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. *Advances in Large Margin Classifiers* **1999**, 61–74.

(12) Zadrozny, B.; Elkan, C. Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. **2001**, 8.

(13) Van der Laan, M. J.; Laan, M.; Robins, J., *Unified methods for censored longitudinal data and causality*; Springer Science & Business Media: 2003.

(14) Almond, D.; Chay, K. Y.; Lee, D. S. The Costs of Low Birth Weight. *The Quarterly Journal of Economics* **2005**, *120*, Publisher: Oxford University Press, 1031–1083.

(15) Niculescu-Mizil, A.; Caruana, R. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, ACM Press: 2005, pp 625–632.

(16) Rubin, D. B. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association* **2005**, *100*, Publisher: Taylor & Francis, 322–331.

(17) Laan, M. J. v. d.; Polley, E. C.; Hubbard, A. E. Super Learner. *Statistical Applications in Genetics and Molecular Biology* **2007**, *6*.

(18) Ionides, E. L. Truncated Importance Sampling. *Journal of Computational and Graphical Statistics* **2008**, *17*, 295–311.

(19) Rahimi, A.; Recht, B. In *Advances in Neural Information Processing Systems*, 2008; Vol. 20.

(20) Rothman, K.; Greenland, S.; Lash, T. Case-control studies, chapter 8. *Modern epidemiology* **2008**, 111–127.

(21) Crump, R. K.; Hotz, V. J.; Imbens, G. W.; Mitnik, O. A. Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **2009**, *96*, 187–199.

(22) 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]* **2009**.

(23) Austin, P. C. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. *Multivariate Behavioral Research* **2011**, *46*, 399–424.

(24) Dudley, J. T.; Deshpande, T.; Butte, A. J. Exploiting drug–disease relationships for computational drug repositioning. *Briefings in bioinformatics* **2011**, *12*, 303–311.

(25) Hill, J. L. Bayesian Nonparametric Modeling for Causal Inference. *Journal of Computational and Graphical Statistics* **2011**, *20*, 217–240.

(26) Howe, C. J.; Cole, S. R.; Westreich, D. J.; Greenland, S.; Napravnik, S.; Eron, J. J. Splines for trend analysis and continuous confounder control. *Epidemiology (Cambridge, Mass.)* **2011**, *22*, 874–875.

(27) Laan, M. J. v. d.; Rose, S., *Targeted Learning*; Springer Series in Statistics, 2011.

(28) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(29) Snowden, J. M.; Rose, S.; Mortimer, K. M. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. *American Journal of Epidemiology* **2011**, *173*, 731–738.

(30) Gretton, A.; Borgwardt, K. M.; Rasch, M. J.; Schölkopf, B.; Smola, A. A kernel two-sample test. *The Journal of Machine Learning Research* **2012**, *13*, 723–773.

(31) Hurle, M. R.; Yang, L.; Xie, Q.; Rajpal, D. K.; Sanseau, P.; Agarwal, P. Computational drug repositioning: from data to therapeutics. *Clinical Pharmacology & Therapeutics* **2013**, *93*, 335–341.

(32) Rolling, C. A.; Yang, Y. Model selection for estimating treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **2014**, *76*, 749–769.

(33) 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* **2015**, *34*, 3661–3679.

(34) Imbens, G. W.; Rubin, D. B., *Causal inference in statistics, social, and biomedical sciences*; Cambridge University Press: 2015.

(35) Swaminathan, A.; Joachims, T. In *International Conference on Machine Learning*, 2015, pp 814–823.

(36) Austin, P. C.; Stuart, E. A. Estimating the effect of treatment on binary outcomes using full matching on the propensity score. *Statistical methods in medical research* **2017**, *26*, 2505–2525.

(37) 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* **2017**, 14.

(38) 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* **2017**.

(39) Nie, X.; Wager, S. Quasi-Oracle Estimation of Heterogeneous Treatment Effects. *Biometrika* **2017**, *108*, 299–319.

(40) Schulam, P.; Saria, S. Reliable Decision Support using Counterfactual Models. *Advances in neural information processing systems* **2017**, *30*.

(41) Schuler, M. S.; Rose, S. Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies. *American Journal of Epidemiology* **2017**, *185*, 65–73.

(42) Shalit, U.; Johansson, F. D.; Sontag, D. In *International Conference on Machine Learning*, 2017, pp 3076–3085.

(43) 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* **2018**, 71.

(44) Daniel, R. M. In *Wiley StatsRef: Statistics Reference Online*; John Wiley & Sons, Ltd: 2018, pp 1–14.

(45) Mooney, S. J.; Pejaver, V. Big data in public health: terminology, machine learning, and privacy. *Annual review of public health* **2018**, *39*, 95.

(46) 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* **2018**, *37*, 1767–1787.

(47) 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]* **2018**.

(48) Shimoni, Y.; Yanover, C.; Karavani, E.; Goldschmnidt, Y. Benchmarking Framework for Performance-Evaluation of Causal Inference Analysis. *arXiv:1802.05046 [cs, stat]* **2018**.

(49) Wager, S.; Athey, S. Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association* **2018**, *113*, 1228–1242.

(50) 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* **2018**, *37*, 3309–3324.

(51) Alaa, A.; Schaar, M. V. D. Validating Causal Inference Models via Influence Functions. *International Conference on Machine Learning* **2019**, 191–201.

(52) Athey, S.; Tibshirani, J.; Wager, S. Generalized random forests. *Annals of Statistics* **2019**, *47*, Publisher: Institute of Mathematical Statistics, 1148–1178.

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

(54) 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* **2019**, *116*, Publisher: National Academy of Sciences Section: PNAS Plus, 4156–4165.

(55) Perperoglou, A.; Sauerbrei, W.; Abrahamowicz, M.; Schmid, M. A review of spline function procedures in R. *BMC Medical Research Methodology* **2019**, *19*, 46.

(56) 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* **2020**, *230*, Publisher: Elsevier, 101–112.e2.

(57) Hernán, M.; Robins, J., *Causal Inference: What If*. 2020.

(58) Jesson, A.; Mindermann, S.; Shalit, U.; Gal, Y. Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models. *Advances in Neural Information Processing Systems* **2020**, *33*, 11637–11649.

(59) Kennedy, E. H. Optimal doubly robust estimation of heterogeneous causal effects. *arXiv preprint arXiv:2004.14497* **2020**.

(60) Poldrack, R. A.; Huckins, G.; Varoquaux, G. Establishment of best practices for evidence for prediction: a review. *JAMA psychiatry* **2020**, *77*, 534–540.

(61) Saito, Y.; Yasui, S. In *International Conference on Machine Learning*, 2020, pp 8398–8407.

(62) Athey, S.; Wager, S. Policy Learning with Observational Data. *Econometrica* **2021**, *89*, 133–161.

(63) Bouthillier, X. et al. Accounting for Variance in Machine Learning Benchmarks. *Proceedings of Machine Learning and Systems* **2021**, *3*, 747–769.

(64) 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* **2021**, 14.

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

(66) Hernán, M. A. Methods of Public Health Research — Strengthening Causal Inference from Observational Data. *New England Journal of Medicine* **2021**, *385*, 1345–1348.

(67) Johansson, F. D.; Shalit, U.; Kallus, N.; Sontag, D. Generalization Bounds and Representation Learning for Estimation of Potential Outcomes and Causal Effects. *arXiv:2001.07426 [cs, stat]* **2021**.

(68) Minderer, M.; Djolonga, J.; Romijnders, R.; Hubis, F.; Zhai, X.; Houlsby, N.; Tran, D.; Lucic, M. Revisiting the Calibration of Modern Neural Networks. *Advances in Neural Information Processing Systems* **2021**, *34*, 15682–15694.

(69) Naimi, A. I.; Mishler, A. E.; Kennedy, E. H. Challenges in Obtaining Valid Causal Effect Estimates with Machine Learning Algorithms. *American Journal of Epidemiology* **2021**.

(70) Perez-Lebel, A.; Morvan, M. L.; Varoquaux, G. Beyond calibration: estimating the grouping loss of modern neural networks. *arXiv preprint arXiv:2210.16315* **2022**.

(71) Varoquaux, G.; Colliot, O. Evaluating machine learning models and their diagnostic value, 2022.

APPENDIX

A VARIABILITY OF ATE ESTIMATION ON ACIC 2016

Figure 2 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[63].

Outcome models are implemented with scikit-learn[28] and the following hyper-parameters:

Outcome Model	Hyper-parameters grid
Boosted Trees (Histogram-based Gradient Boosting)	Learning rate: [0.01, 0.1]
Ridge regression without treatment interaction	Ridge regularization: [0.001, 0.1,]
Ridge regression with treatment interaction	Ridge regularization: [0.001, 0.1,]

TABLE A1 Hyper-parameters grid used for ACIC 2016 ATE variability

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[16]:

Assumption 1 (Unconfoundedness)**.**

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

This condition –also called ignorability– is equivalent to the conditional independence on $e(X)$[4]: $\{Y(0), Y(1)\} \perp\!\!\!\perp A|e(X)$.

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[65], 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[57].

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[58].

C PROOFS: LINKS BETWEEN FEASIBLE AND ORACLE RISKS

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 8 (Population Factual μ-risk)**.** [42]

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

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

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

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.

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

□

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),

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

Proof.

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

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

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

Applying Lemma 1,

$$\tau\text{-risk}(f) \leq 2\left[p_A \mu\text{-risk}_{IPW,1}(w,f) + (1-p_A)\mu\text{-risk}_{IPW,0}(w,f)(w,f)\right] - 2(\sigma^2_{Bayes}(0) + \sigma^2_{Bayes}(1))$$
$$= 2\mu\text{-risk}_{IPW}(w,f) - 2(\sigma^2_{Bayes}(0) + \sigma^2_{Bayes}(1))$$

□

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

Proposition 2 (R-risk as reweighted τ-risk)**.** *Proof.* We consider the R decomposition:[8],

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

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

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

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

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

The second term is null since, $\mathbb{E}[\varepsilon(x,a)|X,A] = 0$.
The third term corresponds to the modulated residuals 4 : $\tilde{\sigma}^2_B(0) + \tilde{\sigma}^2_B(1)$

□

D MEASURING OVERLAP

Motivation of NTV

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

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

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

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

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

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 (cf. Figure D3),
- Verifying that setups with penalized overlap from ACIC 2016 have a higher total variation distance than unpenalized setups (cf. Figure D2).
- Verifying that the Inverse Propensity Weights extrema (the inverse of the ν overlap constant appearing in the overlap Assumption 2) augments with NTV for Caussim, ACIC 2016 and Twins (cf. Figure D4). 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 D1 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.

E EXPERIMENTS

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

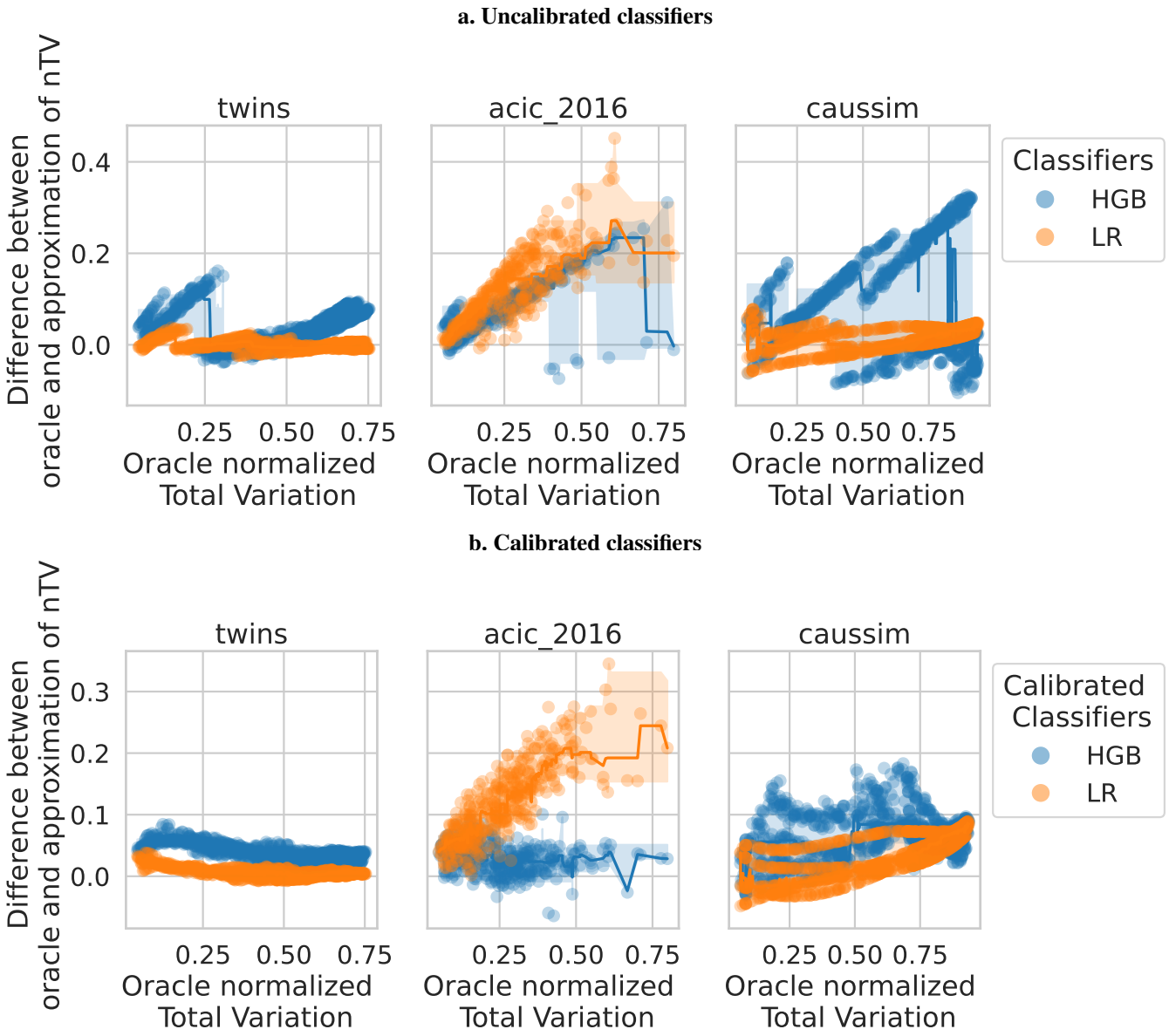


FIGURE D1 a) Without calibration, estimation of NTV is not trivial even for boosting models. b) Calibrated classifiers are able to recover the true Normalized Total Variation for all datasets where it is available.

be seen as a stochastic approximation of the full data kernel[19]). We carefully control the amount of overlap between treated and control populations, a crucial assumption for causal inference.

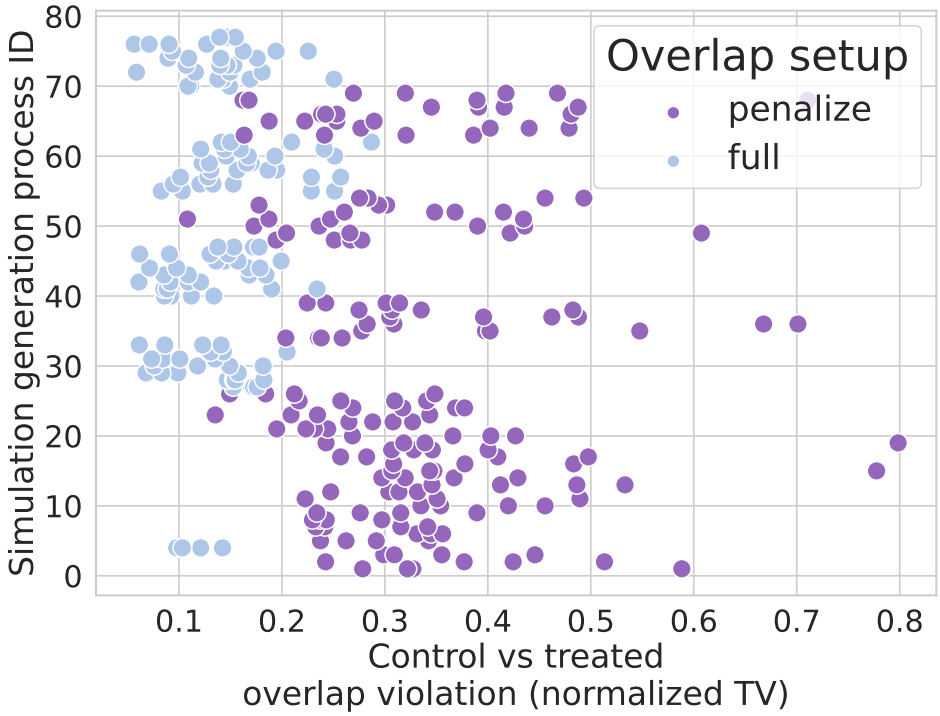
- The raw features for both populations are drawn from a mixture of Gaussians: $\mathbb{P}(X) = p_A\mathbb{P}(X|A=1)+(1-p_A)\mathbb{P}(X|A=0)$ where $\mathbb{P}(x|A=a)$ is a rotated Gaussian:

$$\mathbb{P}(x|A=a) = W \cdot \mathcal{N}\left(\begin{bmatrix} (1-2a)\theta \\ 0 \end{bmatrix}; \begin{bmatrix} \sigma_0 & 0 \\ 0 & \sigma_1 \end{bmatrix} \right) \tag{E1}$$

  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.

FIGURE D2 NTV recovers well the overlap settings described in the ACIC paper[53]



- A basis expansion of the raw features increases the problem dimension. Using Radial Basis Function (RBF) Nystroem transformation [7], 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 $z(x)$.

  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:

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

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

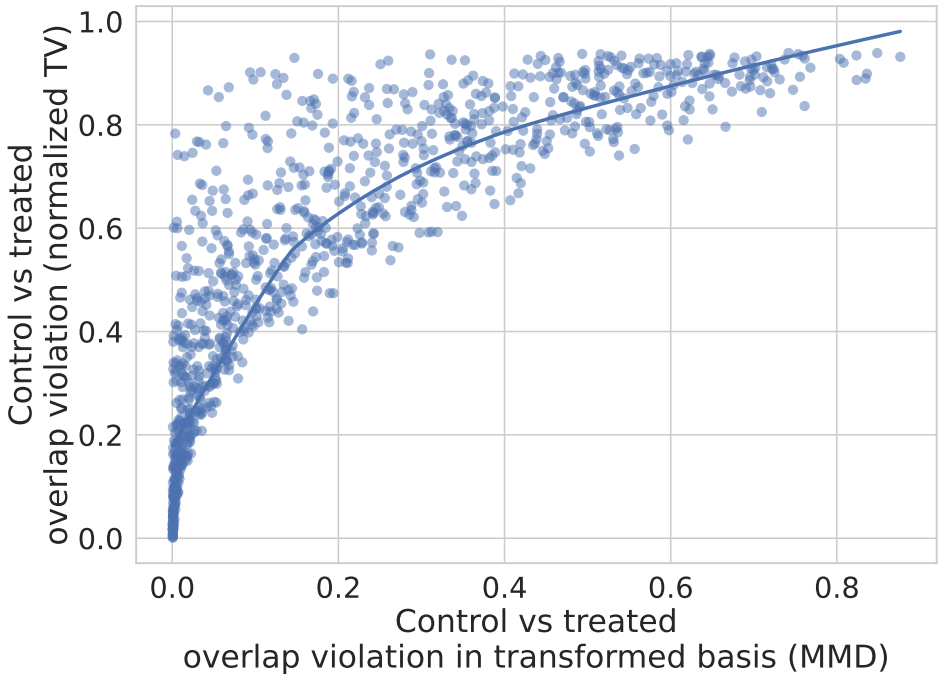
E.2 Model selection procedures

Nuisances estimation

The nuisances are estimated with a stacked regressor inspired by the Super Learner framework,[17]). The hyper-parameters are optimized with a random search with following search grid detailed in Table E2. All implementations come from scikit-learn[28].

[7] We use the Sklearn implementation,[28]

FIGURE D3 Good correlation between overlap measured as normalized Total Variation and Maximum Mean Discrepancy (200 sampled Caussim datasets)



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

TABLE E2 Hyper-parameters grid used for nuisance models

Similarity of results between the three-sets and the chosen two-sets procedure

Figure E5 shows that very few difference appears between a two-sets procedure – nuisances fitted on the same train set as the candidates–, and a three-sets procedure –nuisances fitted on a separated validation set of the same size as the train set.

E.3 Additional Results

Results measured with the semi-oracle *R*-risk Kendall's as reference

We inspected if the observed variability between dataset instances is due to inter-experiments variability or to intra-experiment variability. Is the *R*-risk systematically better than mu-risk to select the best model among the family of candidates ? Figure E6 shows the differences between every metrics and the semi-oracle R-risk Kendall's. The difference κ is consistently greater than zero for the four datasets. It is significant at 5% for all overlap setups only for Caussim and Twins. This confirms than R-risk is better in every experimental setups.

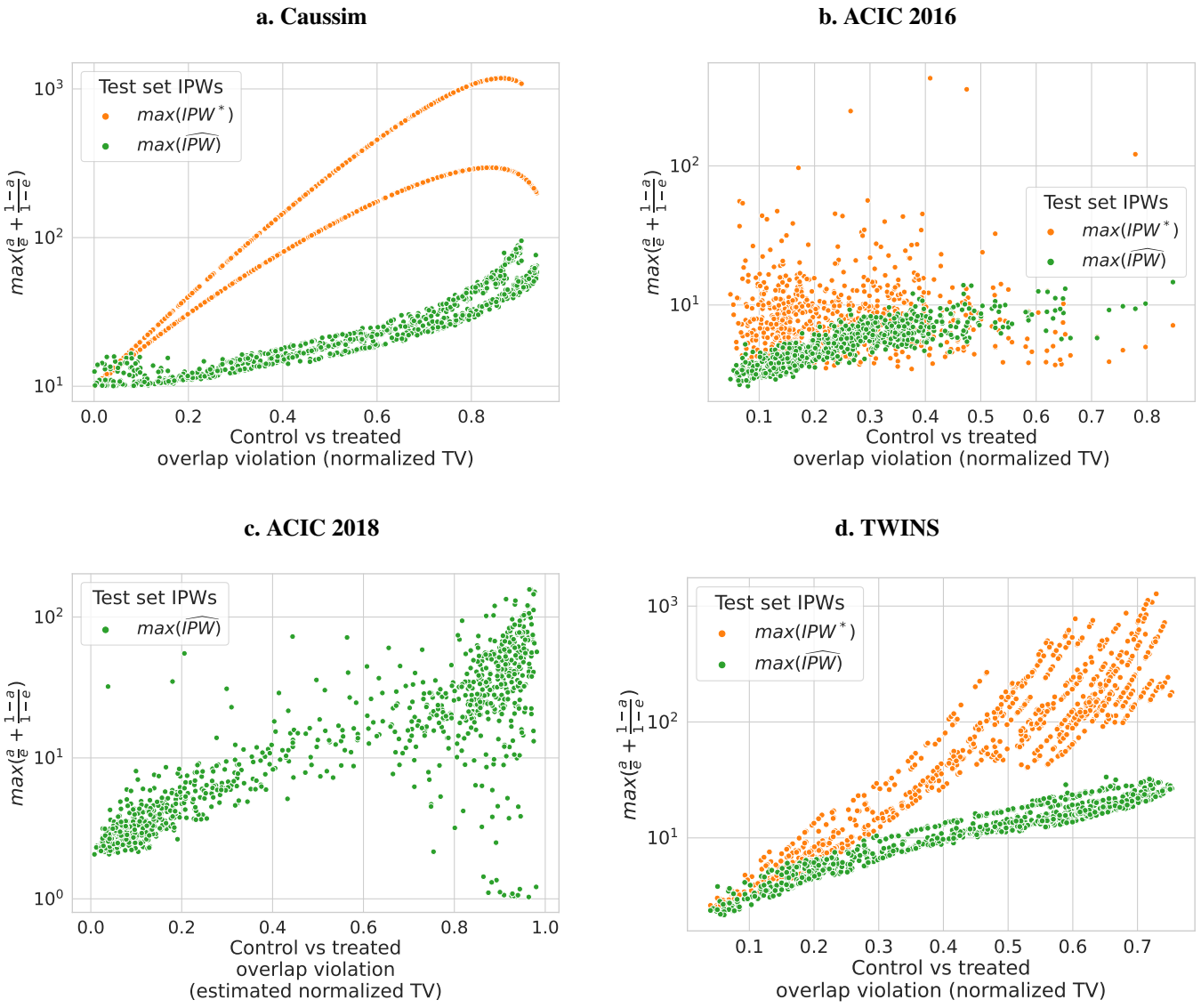


FIGURE D4 Maximal value of Inverse Propensity Weights increases exponentially with the overlap as measure by Normalized Total Variation.

Results measured as ranking aggreement with the oracle tau-risk

Figure E7 shows the percentage of experiments for which each metric selects the same best model as the oracle τ-risk. We plot this percentage of agreement as a function of decreasing overlap (by increasing normalized Total Variation).

Results measured as distance to the oracle tau-risk

Figure E8 reports the results between metrics as the normalized distance between the estimator selected by the oracle τ-risk and the estimator selected by each causal metric.

We recover the ordering of the oracle $\widehat{R\text{-risk}}_{IS^{2*}}$ is the best performing selection metric, especially for poor overlap settings. Then, $\widehat{R\text{-risk}}^{*}$ is more efficient than both oracle and feasible versions of $\widehat{\mu\text{-risk}}_{IPW}$ which are themselves an order of magnitude better than the classical $\widehat{\mu\text{-risk}}$. Importantly, this gap is small in strong overlap settings and grows rapidly with the lack of overlap.

ACIC 2016, effect of misspecified nuisance models

Figure E9 shows the effect of misspecification for the nuisance models. On Caussim, we compare non-linear nuisance estimators (stacked boosting and linear estimators) to linear (misspecified) estimators of the nuisance $(\check{e}, \check{m})$.

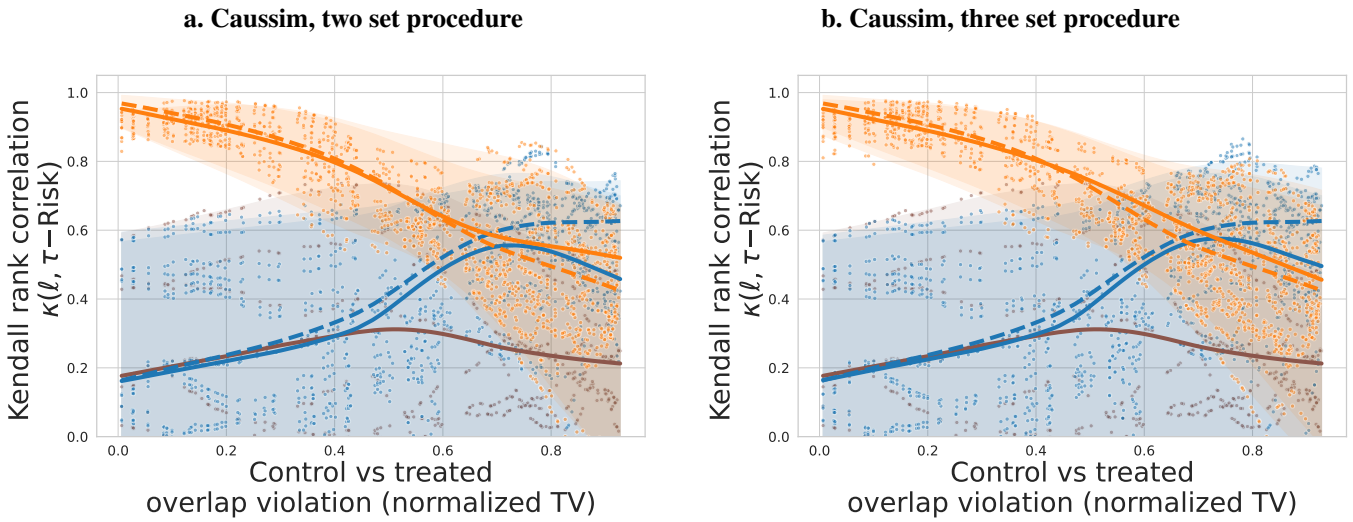


FIGURE E5 Simulation, 900 instances (3 seeds, 3 treatment ratios, 100 overlap parameters). Results are similar between the two-set procedure **a.** and the three set procedure **b.**.

Misspecified linear estimators for the nuisances have a big impact on feasible metrics for the $\widehat{R\text{-risk}}$ or $\widehat{R\text{-risk}}_{IS2}$. This suggests that (e, m) quantities should be estimated with care if using complex causal metrics for causal estimator selection.

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 E10 illustrate the dependence of the results on the generation process for caussim simulations. This is the same kind of plots and experiments as in Figure 5, but with only 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 $R\text{−}risk$, but vary strongly for μ-risk and $\mu\text{-risk}_{IPW}$.

Simulations: naive reweighting of the R-risk

Applying a naive reweighting, $w(x, a) = \frac{1}{e(x)\left(1-e(x)\right)}$ to the R-risk to recover the τ-risk in the first part of 2 makes the residuals explode in case of noise as shown in Figure E11.

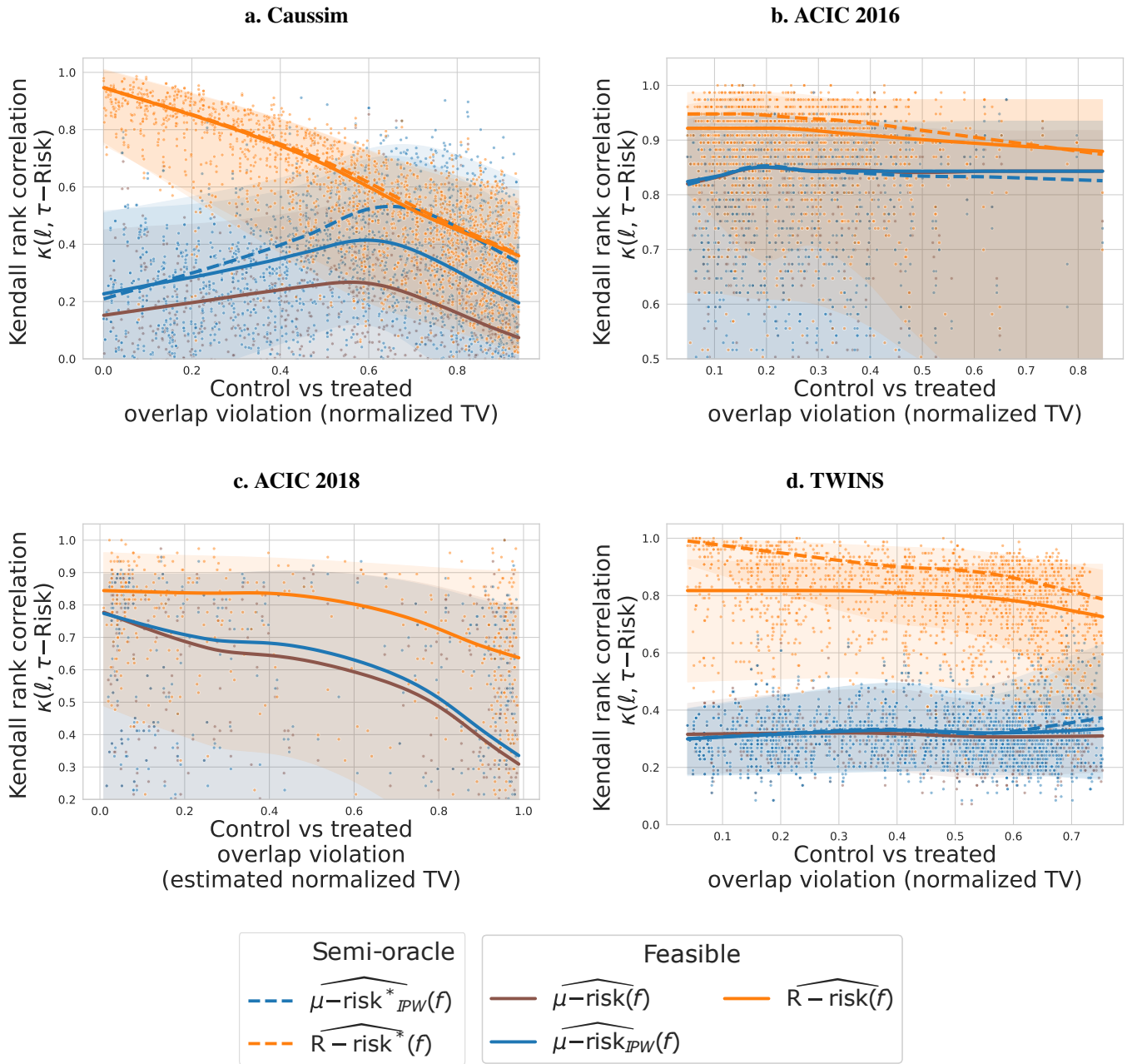


FIGURE E6 Agreement with τ-risk ranking of methods : The reference is the semi-oracle R-risk Kendall's coefficient. Dotted lines are oracle metrics 50% lowess quantiles and plain lines are feasible metrics 50% lowess quantiles. The 5% and 95% confidence intervals are indicated by the transparent bands.

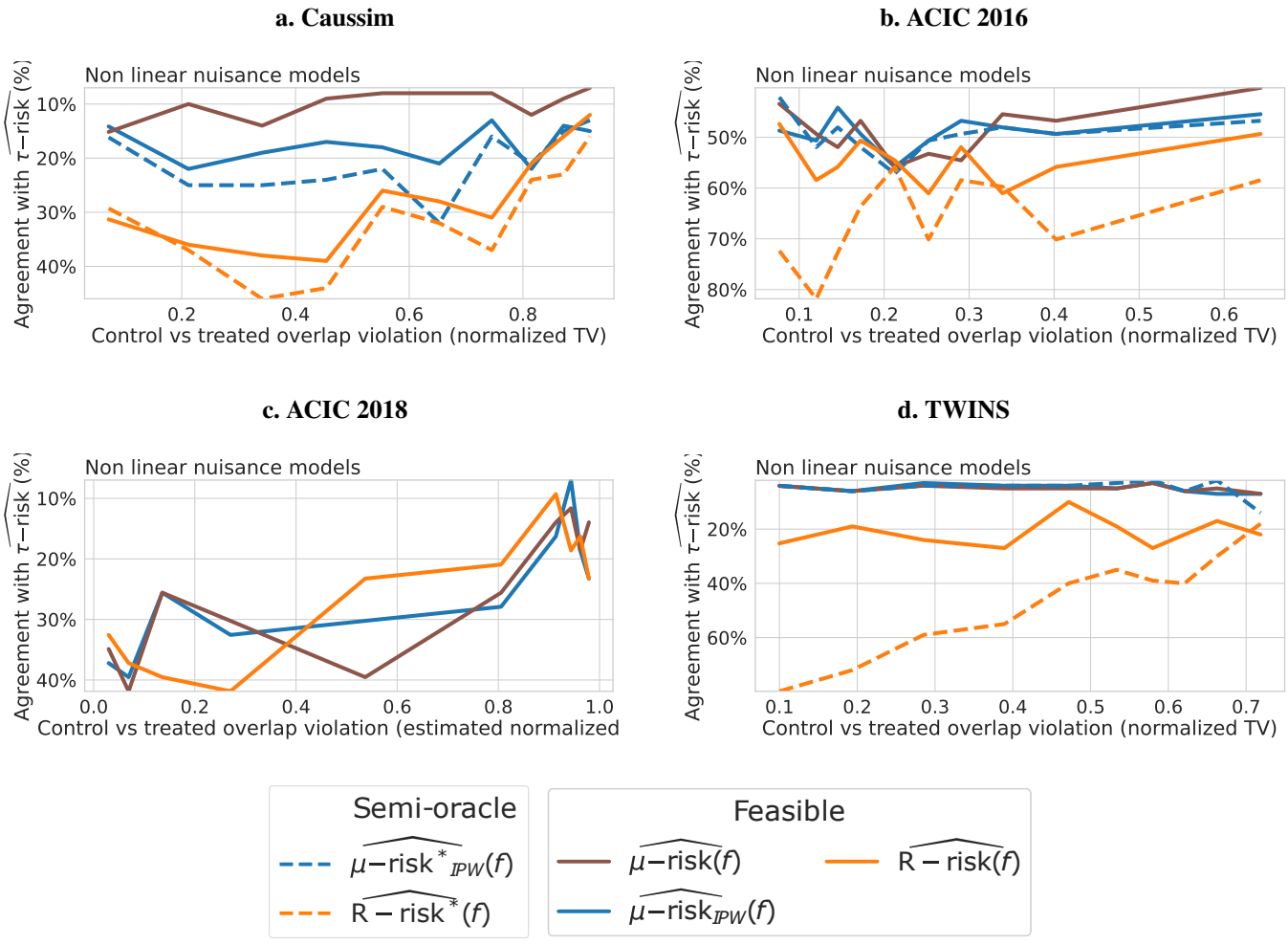


FIGURE E7 Agreement with τ-risk decreases with overlap violation.

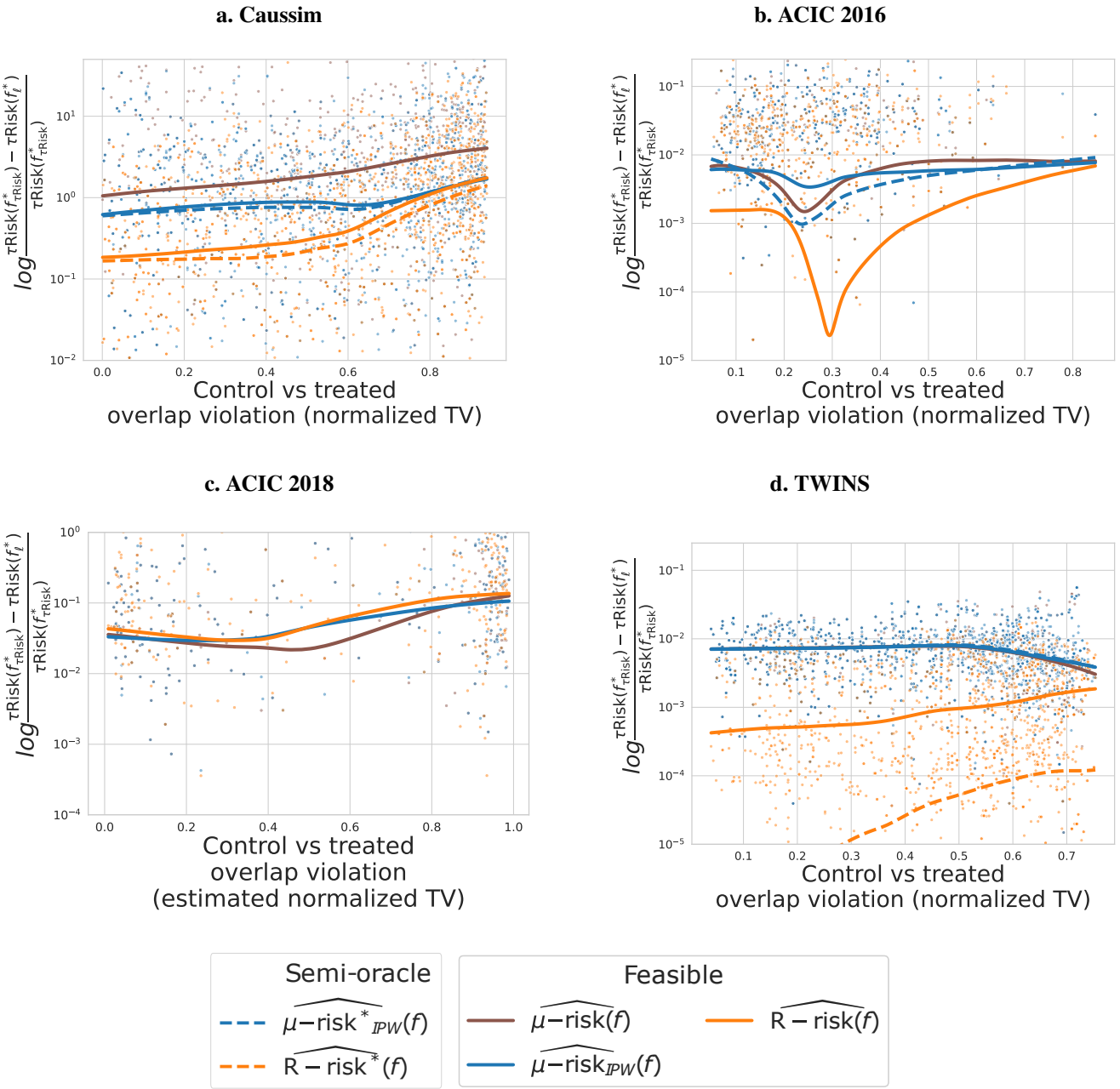


FIGURE E8 Metric performances by normalized tau-risk distance to the best method selected with τ-risk. All nuisances are learned with the same estimator stacking gradient boosting and ridge regression. Doted and plain lines corresponds to 60% lowess quantile estimates.

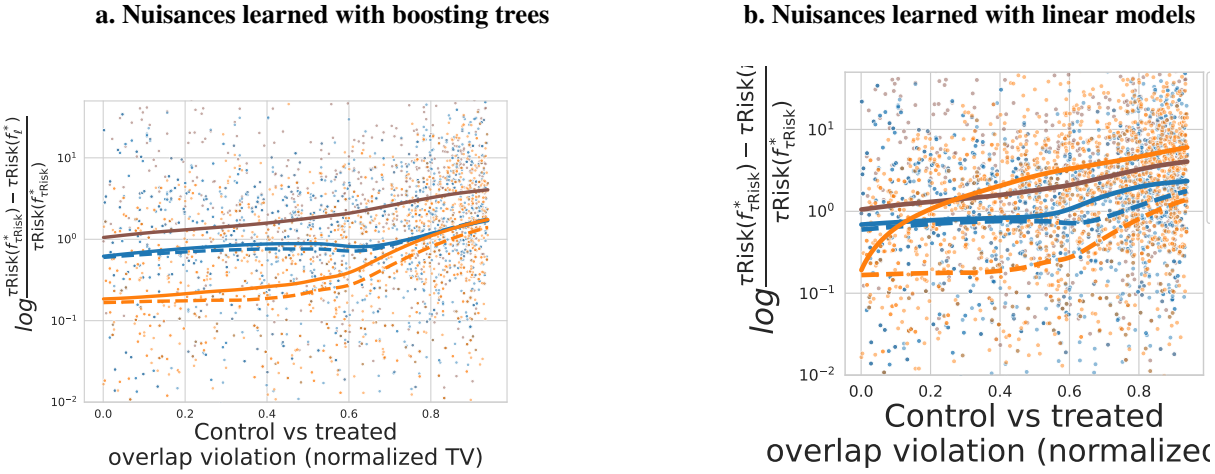


FIGURE E9 a. Candidate estimators selected with oracles are recovering the best normalized τ-risk. Consistent with the theory, the best causal metric is the R-risk*. However, if the nuisances are not well specified as in **b.**, simple IPW reweighting of Mean Squared Error achieves reasonable results across all overlap settings and feasible R-risks fail. Lines are lowess estimates of the 0.5 quantiles for each causal selection metric across all 1000 experimental setups.

FIGURE E10 Kendall correlation coefficients for each causal metric. Each (color, shape) pair indicates a different (treatment ratio, seed) of the generation process.

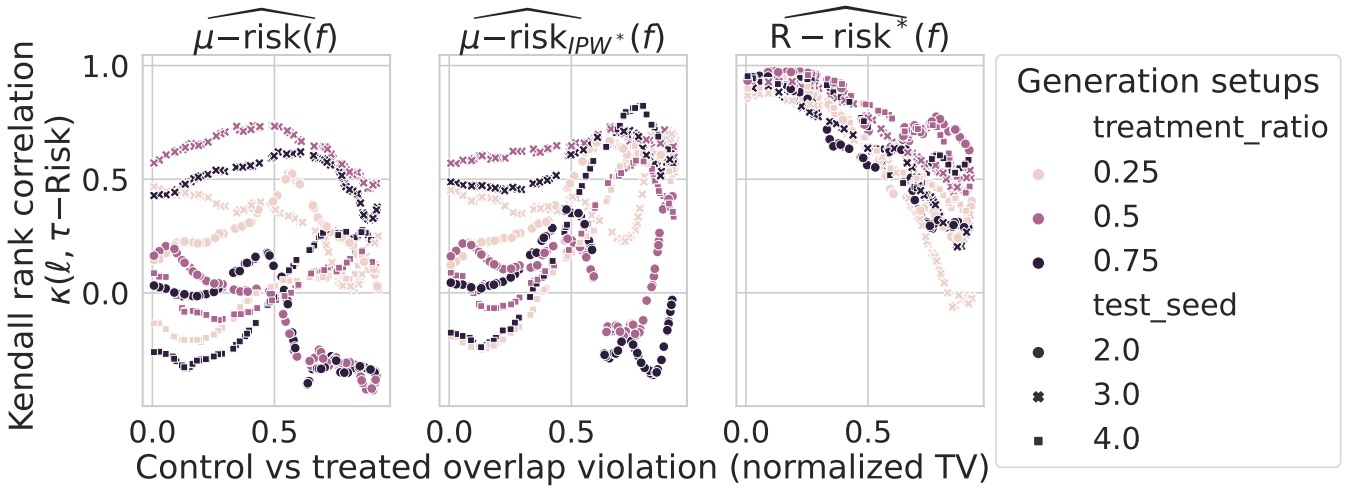


FIGURE E11 Caussim simulations (500 repetitions): R-risk$_{IPW^*_{naive}}$ (in green) is exploding