

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

Matthieu Doutréline, PhD<sup>1,2,\*</sup> and Gaël Varoquaux, PhD<sup>1</sup>

<sup>1</sup>Inria, Soda, Saclay, France

<sup>2</sup>Mission Data, Haute Autorité de Santé, Saint-Denis, France

\*Corresponding author: matt.dout@gmail.com

January 10, 2024

## Abstract

**Objective:** We investigate which procedure selects the predictive model most trustworthy to reason on the effect of an intervention and support decision making.

**Methods:** We study such a variety of model selection procedures in practical settings: finite samples settings and without theoretical assumption of *well-specified* models. Beyond standard cross-validation or internal validation procedures, we also study elaborate causal risks. These build proxies of the causal error using “nuisance” re-weighting to compute it on the observed data. We evaluate whether empirically estimated nuisances, which are necessarily noisy, add noise to model selection. We compare different metrics for causal model selection in an extensive empirical study based on a simulation and three healthcare datasets based on real covariates.

**Results:** Among all metrics, the mean squared error, classically used to evaluate predictive models, is worse. Re-weighting it with propensity score does not bring much improvements. The  $R$ -risk, which uses as nuisances a model of mean outcome and propensity scores, leads to the best performances. Nuisance corrections are best estimated with flexible estimators such as a super learner.

**Conclusions:** When predictive models are used to reason on the effect of an intervention, they must be evaluated with different procedures than standard predictive settings; using the  $R$ -risk from causal inference.

**Keywords** Model Selection, Predictive model, Treatment Effect, G-formula, Machine Learning,

# Introduction

## Extending prediction to prescription needs causality

Prediction models have long been used in medicine, as with risk score or prognostic models<sup>47,82</sup>. While these have historically been simple models on simple data, this is changing with progress in machine learning and richer medical data<sup>9,61</sup>. Health predictions can now integrate medical images<sup>36,93,91,73,50</sup>, patient records<sup>46,17,77</sup> or clinical notes<sup>29,88,79</sup>. Complex data is difficult to control and model, but these models are validated by verifying the accuracy of the prediction on left-out data<sup>3,58,86</sup>. Crucial to the clinical adoption of a model predicting a health outcome is that it “can support decisions about patient care”<sup>90</sup>. Precision medicine is about guiding decisions: *eg* will an individual benefit from an intervention such as surgery<sup>20</sup>? An estimate of the effect of the treatment can be obtained by contrasting model predictions with and without the treatment, but statistical validity requires causal inference<sup>78,80,10</sup>.

Indeed, concluding on the effect of a treatment is a difficult causal-inference task, as it can be easily compromised by confounding: spurious associations between treatment allocation and baseline health, *e.g.* only prescribing a drug to mild cases<sup>26,85</sup>. Predictive modeling bridges to causal inference theory under the name of *outcome models* (or G-computation, G-formula<sup>63</sup>, Q-model<sup>78</sup>, conditional mean regression<sup>89</sup>). Medical statistics and epidemiology have mostly used other causal-inference methods, modeling treatment assignment with propensity scores<sup>66,7,12,23</sup>. Outcome modeling brings the benefit of going 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<sup>83,41,28</sup>.

Outcome-modeling methods, even when specifically designed for causal inference, are numerous: Bayesian Additive Regression Trees<sup>27</sup>, Targeted Maximum Likelihood Estimation<sup>40,71</sup>, causal boosting<sup>59</sup>, causal multivariate adaptive regression splines<sup>59</sup>, random forests<sup>87,5</sup>, Meta-learners<sup>37</sup>, R-learners<sup>52</sup>, Doubly robust estimation<sup>14</sup>... 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<sup>19,18</sup> (illustration in Appendix A).

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

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

We first illustrate the problem of causal model selection. Then we anchor causal model selection in the *potential outcome* framework and details the causal risks and model-selection procedure. We then rewrite the so-called  $R$ -risk as a reweighted version of mean squared difference between the true and estimated individualized treatment effect. Finally, we conduct a thorough empirical study comparing the different metrics on diverse datasets, using a family of simulations and real health data, going beyond prior work limited to specific simulation settings<sup>70,1</sup> (Appendix B).

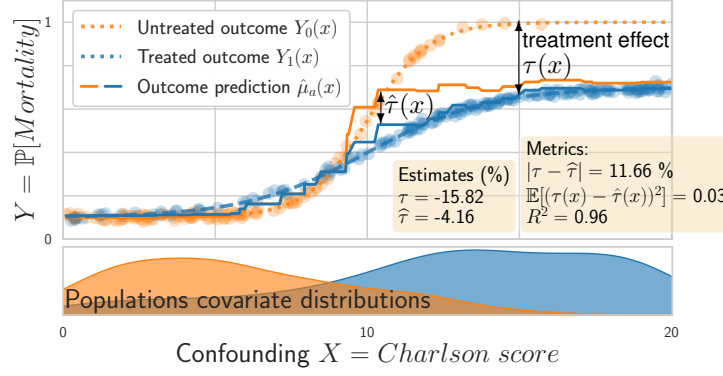
### **Illustration: the best predictor may not estimate best causal effects**

Using a predictor to reason on causal effects relies on contrasting the prediction of the outcome for a given individual with and without the treatment. Given various predictors of the outcome, which one should we use? Standard predictive modeling or machine-learning practice selects the predictor that minimizes the expected error on the outcome<sup>58,86</sup>. However, this predictor may not be the best model to reason about causal effects of an intervention as Figure 1 illustrates. Consider the probability  $Y$  of an undesirable outcome (*e.g.* death), a binary treatment  $A \in \{0, 1\}$ , and a covariate  $X \in \mathbb{R}$  summarizing the patient health status (*e.g.* the Charlson index<sup>13</sup>). We simulate a treatment beneficial (decreases mortality) for patients with high Charlson scores (bad health status) but with little effect for patients in good condition (low Charlson scores).

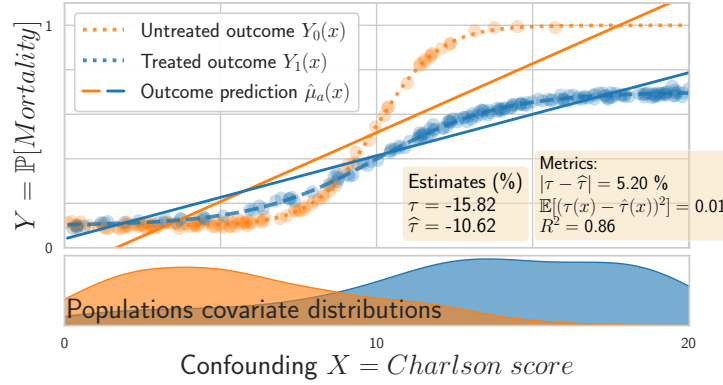
Figure 1a shows a random forest predictor with a counter-intuitive behavior: it predicts well on average the outcome (as measured by a regression  $R^2$  score) but perform poorly to estimate causal quantities: the average treatment effect  $\tau$  (as visible via the error  $|\tau - \hat{\tau}|$ ) or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

The problem is that causal estimation requires controlling an error on both treated and non-treated outcome for the same individual: the observed outcome, and the non-observed *counterfactual* one. The linear model is misspecified—the outcome functions are not linear—, leading to poor  $R^2$ ; but it interpolates

a) Random forest, good average prediction but bad causal inference



b) Linear model, worse average prediction but better causal inference



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

Selecting the predictor with the smallest error to the individual treatment effect  $\mathbb{E}[(\tau(x) - \hat{\tau}(x))^2]$  –the  $\tau$ -risk, eq. 9 – 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.

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.

## Methods

### Neyman-Rubin Potential Outcomes framework

**Settings** The Neyman-Rubin Potential Outcomes framework<sup>49,31</sup> enables statistical reasoning on causal treatment effects: Given an outcome  $Y \in \mathbb{R}$  (e.g.mortality risk or hospitalization length), function of a

binary treatment  $A \in \mathcal{A} = \{0, 1\}$  (e.g. a medical procedure), and baseline covariates  $X \in \mathcal{X} \subset \mathbb{R}^d$ , we observe the factual distribution,  $O = (Y(A), X, A) \sim \mathcal{D} = \mathbb{P}(y, x, a)$ . However, we want to model the existence of potential observations (unobserved ie. counterfactual) that correspond to a different treatment. Thus we want quantities on the counterfactual distribution  $O^* = (Y(1), Y(0), X, A) \sim \mathcal{D}^* = \mathbb{P}(y(1), y(0), x, a)$ .

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

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

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

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

**Causal assumptions** A given data needs to meet a few assumptions to enable identifying causal estimands<sup>67</sup>. The usual *strong ignorability* assumptions are (details in [C](#)): 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*.

**Estimating treatment effects with outcome models – g-computation<sup>62</sup>** 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 estimated from observed data: the consistency [3](#) and unconfoundedness [1](#) assumptions imply the following equality:

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

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. 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<sup>78</sup>):

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

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*:

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

Given a sample of data and the oracle response functions  $\mu_0, \mu_1$ , 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  $\tau$ ,  $\mathbb{E}[\mu_1(x_i) - \mu_0(x_i)]$ , which involves an expectation taken on the full distribution  $\mathcal{D}$ , which is observable but requires infinite data. For each estimator  $\ell$  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<sup>64</sup> plays an important role: introducing two quantities, the conditional mean outcome and the probability to be treated (known as propensity score<sup>66</sup>):

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

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

the outcome can be written

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

$m$  and  $e$  are often called *nuisances*<sup>14</sup>; they are unknown.

**Table 1:** Review of causal risks — The  $R$ -risk\* is called  $\tau$ -risk $_R$  in Schuler et al. [70].

Risk	Equation	Reference
$\tau$ -risk = $\text{MSE}(\tau(X), \tau_f(X))$	$\mathbb{E}_{X \sim p(X)}[(\tau(X) - \hat{\tau}_f(X))^2]$	Eq. 9 <sup>27</sup>
$\mu$ -risk = $\text{MSE}(Y, f(X))$	$\mathbb{E}_{(Y, X, A) \sim \mathcal{D}}[(Y - f(X; A))^2]$	Def. 1 <sup>70</sup>
$\mu$ -risk $^*_{IPW}$	$\mathbb{E}_{(Y, X, A) \sim \mathcal{D}}\left[\left(\frac{A}{e(X)} + \frac{1-A}{1-e(X)}\right)(Y - f(X; A))^2\right]$	Def. 2 <sup>38</sup>
$\tau$ -risk $^*_{IPW}$	$\mathbb{E}_{(Y, X, A) \sim \mathcal{D}}\left[\left(Y\left(\frac{A}{e(X)} - \frac{1-A}{1-e(X)}\right) - \hat{\tau}_f(X)\right)^2\right]$	Def. 3 <sup>87</sup>
$U$ -risk*	$\mathbb{E}_{(Y, X, A) \sim \mathcal{D}}\left[\left(\frac{Y-m(X)}{A-e(X)} - \hat{\tau}_f(X)\right)^2\right]$	Def. 4 <sup>52</sup>
$R$ -risk*	$\mathbb{E}_{(Y, X, A) \sim \mathcal{D}}\left[\left((Y - m(X)) - (A - e(X))\hat{\tau}_f(X)\right)^2\right]$	Def. 5 <sup>52</sup>

## 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  $f : \mathcal{X} \times \mathcal{A} \rightarrow \mathcal{Y}$ —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  $\mathcal{F} = \{f : \mathcal{X} \times \mathcal{A} \rightarrow \mathcal{Y}\}$  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  $\ell$ :

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

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

**The  $\tau$ -risk: an oracle error risk** As we would like to target the CATE, the following evaluation risk is natural (also called PEHE<sup>69,27</sup>):

$$\tau\text{-risk}(f) \stackrel{\text{def}}{=} \mathbb{E}_{X \sim p(X)}[(\tau(X) - \hat{\tau}_f(X))^2] \quad (9)$$

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** Table 1 lists *feasible* risks (Detailed in Appendix D), based on the prediction error of the outcome model and *observable* quantities. These observable, called nuisances are  $e$ —propensity score, eq 6— and  $m$ —conditional mean outcome, eq 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.

## Estimation and model selection procedure

Causal model selection (eq 8) may involve estimating various quantities from the observed data: the outcome model  $f$ , its induced risk as introduced in the previous section, and possibly nuisances required by the risk. Given a dataset with  $N$  samples, we split out a train and a test sets  $(\mathcal{T}, \mathcal{S})$ . We fit each candidate estimator  $f \in \mathcal{F}$  on  $\mathcal{T}$ . We also fit the nuisance models  $(\check{e}, \check{m})$  on the train set  $\mathcal{T}$ , setting hyperparameters by a nested cross-validation before fitting the nuisance estimators with these parameters on the full train set. Causal quantities are then computed by applying the fitted candidates estimators  $f \in \mathcal{F}$  on the test set  $\mathcal{S}$ . Finally, we compute the model-selection metrics for each candidate model on the test set. This procedure is described in Algorithm 1 and Figure 2.

---

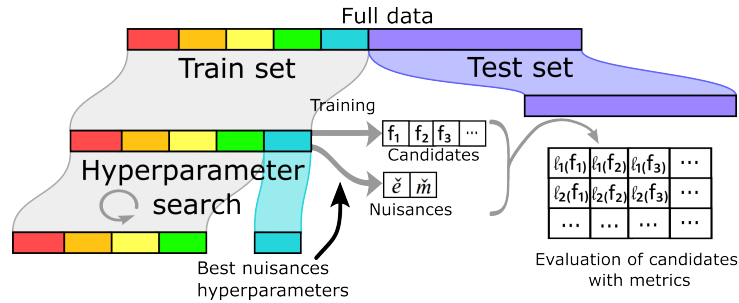
### Algorithm 1 Model selection procedure

---

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

---

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  $(\hat{f}(x, 1), \hat{f}(x, 0))$  and evaluate the estimator storing the metric value:  $\ell(f, \mathcal{S})$  – possibly function of  $\check{e}$  and  $\check{m}$
- 



**Figure 2:** Estimation procedure for causal model selection.

## R-risk as reweighted oracle metric

The  $R$ -risk can be rewritten as a rebalanced  $\tau$ -risk.

This rewriting involves reweighted residuals: for each potential outcome,  $a \in \{0; 1\}$ , the variance conditionally on  $x$  is<sup>72</sup>:

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



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.

**Proposition 1 (*R-risk as reweighted  $\tau$ -risk*)** *Given an outcome model  $f$ , its  $R$ -risk appears as weighted version of its  $\tau$ -risk (Proof in E.1):*

$$\begin{aligned} R\text{-risk}^*(f) &= \int_{\mathcal{X}} e(x)(1 - e(x)) (\tau(x) - \tau_f(x))^2 p(x) dx \\ &\quad + \tilde{\sigma}_B^2(1) + \tilde{\sigma}_B^2(0) \end{aligned} \tag{10}$$

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  $e(x)(1 - e(x))$  are close to  $\frac{1}{4}$ , hence the  $R$ -risk is close to the desired gold-standard  $\tau$ -risk. For randomized control trials, this weight is constant making the  $R$ -risk particularly suited for exploring heterogeneity (Appendix E.2)

## 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 <sup>1</sup>.

### 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<sup>70,1</sup>, we use many generative processes, which is needed to reach general conclusions (Appendix G.4).

We generate the response functions using random bases extension, a common method in biostatistics, *e.g.* functional regression with splines<sup>30,56</sup>. By allowing the function to vary at specific knots, we control the complexity of the non-linear outcome models. We use random approximation of Radial Basis Function (RBF) kernels<sup>60</sup> to generate the outcome and treatment functions. RBF use the same process as polynomial splines but replace polynomial by Gaussian kernels. Unlike polynomial, Gaussian kernels

---

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

have decreasing influences in the input space. This avoids unrealistic divergences of the functions at the ends of the feature space. We generate 1 000 datasets based on these functions, with random overlap parameters. Example shown in Figure S6 and details in G.1.

**Family of candidate estimators** We test model selection across different candidate estimators that approximate imperfectly the data-generating process. To build such estimators, we first use a RBF expansion similar to that used for data generation. We choose two random knots and transform the raw data features with a Gaussian kernel. This step is referred as the featurization. Then, we fit a linear regression on these transformed features. We consider two ways of combining these steps for outcome model; we use common nomenclature<sup>37,74</sup> 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 treated and controls, then two separate different regressors are fitted on this shared representation.
- TLearner: Two completely different learners for each population, hence separate feature representations and regressors.

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

## Semi-simulated datasets

**Datasets** We also use three semi-simulated data adding a known synthetic causal effect to real –non synthetic– healthcare covariate. ACIC 2016<sup>18</sup> is based on the Collaborative Perinatal Project<sup>53</sup>, a RCT studying infants’ developmental disorders containing 4,802 individuals and 55 features. We used 770 dataset instances: 10 random seeds for each of the 77 simulated settings for the treatment and outcomes. ACIC 2018<sup>76</sup> simulated treatment and outcomes for the Linked Births and Infant Deaths Database (LBIDD)<sup>44</sup> with  $D = 177$  covariates. We used all 432 datasets of size  $N = 5\,000$ . Twins<sup>43</sup> is an augmentation of real data on twin births and mortality rates<sup>2</sup>. There are  $N = 11\,984$  samples, and  $D = 50$

covariates for which we simulated 1,000 different treatment allocations. Appendix G.2 gives datasets details.

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

**Nuisance estimators** Drawing from the TMLE literature that uses combination of flexible machine learning methods<sup>71</sup>, we model the nuisances  $\check{e}$  (respectively  $\check{m}$ ) with a meta-learner: a stacked estimator of ridge and boosting classifiers (respectively regressions) (hyperparameter selection in Appendix G.3).

## Measuring overlap between treated and non treated

Good overlap between treated and control population is crucial for causal inference (Assumption 2). We introduce the Normalized Total Variation (NTV), a divergence based on the propensity score summarizing the overlap between both populations (Appendix F).

## Results: factors driving good model selection

**The  $R$ -risk is the best metric** Figure 3 shows the agreement between the ideal ranking of outcome models given the oracle  $\tau$ -risk and the different feasible causal metrics. We measure this agreement with relative<sup>3</sup> Kendall tau  $\kappa$  (eq. 19)<sup>35</sup>. Given the importance of overlap in how well metrics approximate the oracle  $\tau$ -risk, we separate strong and weak overlap.

Among all metrics, the classical mean squared error (ie. factual  $\mu$ -risk) is worse and reweighting it with propensity score ( $\mu$ -risk<sub>IPW</sub>) 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  $\tau$ -risk only 1% higher than the best possible candidate for strong overlap

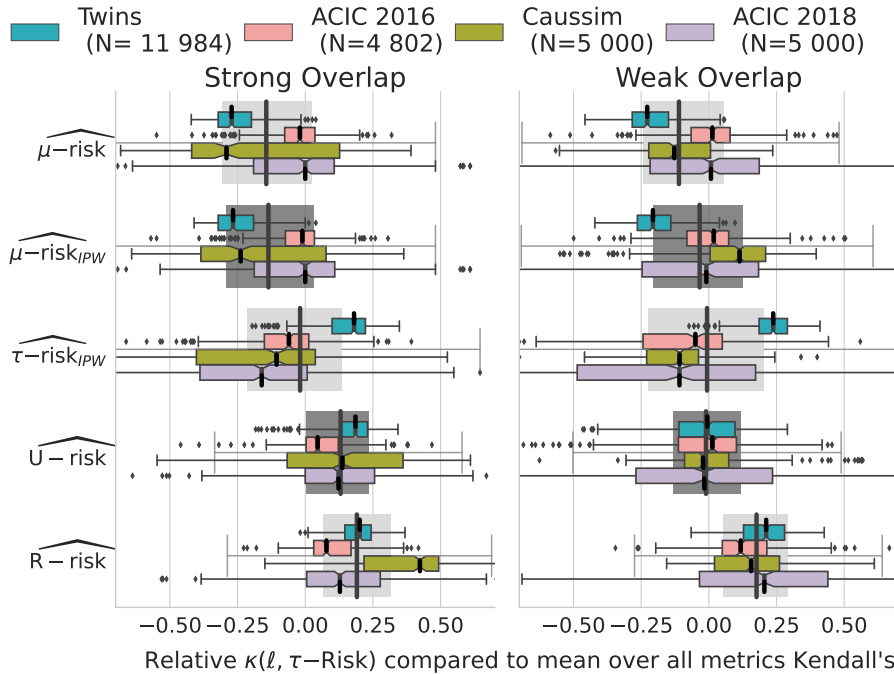
<sup>2</sup>Scikit-learn regressor, `HistGradientBoostingRegressor`, and classifier, `HistGradientBoostingClassifier`.

<sup>3</sup>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:  $\text{Relative } \kappa(\ell, \tau\text{-risk}) = \kappa(\ell, \tau\text{-risk}) - \text{mean}_{\ell}(\kappa(\ell, \tau\text{-risk}))$

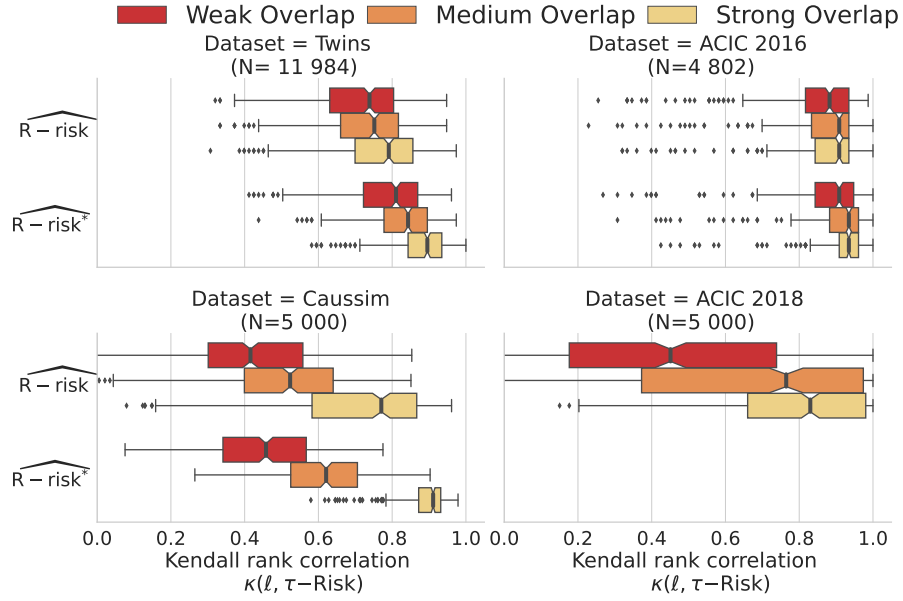
on Caussim, but selecting with the  $\mu$ -risk or  $\mu$ -risk<sub>IPW</sub> –as per machine-learning practice– leads to 10% excess risk and using  $\tau$ -risk<sub>IPW</sub> –as in some causal-inference methods<sup>4,25</sup>–leads to 100% excess risk (Figure S9). Across datasets, the  $R$ -risk consistently decreases the risk compared to the  $\mu$ -risk: from 0.1% to 1% on ACIC2016, 1% from to 20% on ACIC2018, and 0.05% from 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 4). The absolute Kendall’s coefficient correlation with  $\tau$ -risk drops from 0.9 (excellent agreement with oracle selection) to 0.6 on both Caussim and ACIC 2018 (G.4).

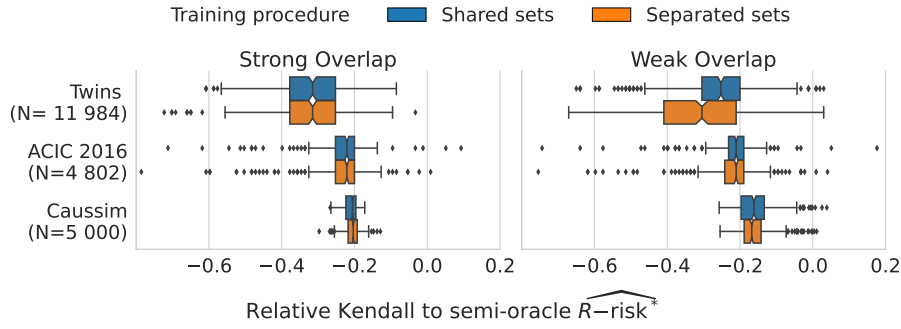
**Nuisances can be estimated on the same data as outcome models** Using the train set  $\mathcal{T}$  both to fit the candidate estimator and the nuisance estimates is a form of double dipping which can lead errors in nuisances correlated to that of outcome models<sup>52</sup>. In theory, these correlations 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 5 shows no substantial difference between a procedure with a separated nuisance set and the simpler shared nuisance-candidate set procedure.



**Figure 3: The  $R$ -risk is the best metric:** Relative Kendall’s  $\tau$  agreement with  $\tau$ -risk. Strong and Weak overlap correspond to the first and last tertiles of the overlap distribution measured with Normalized Total Variation eq. 16. G.4 presents the same results by adding semi-oracle risks in Figure S7, measured with absolute Kendall’s in Figure S8 and with  $\tau$ -risk gains in Figure S9. Table S3 gives median and IQR of the relative Kendall.



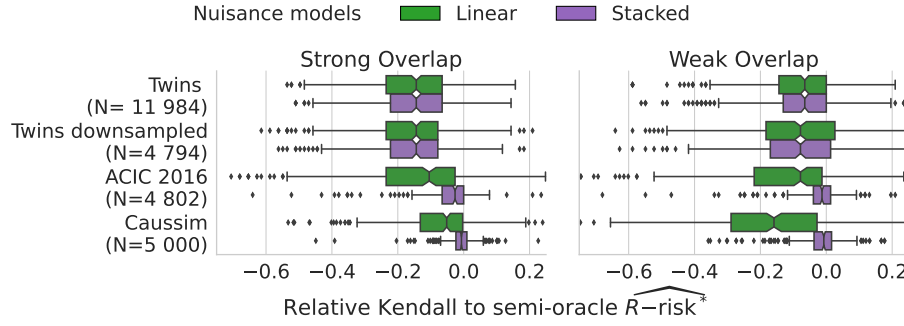
**Figure 4: Model selection is harder for low population overlap:** Kendall's  $\tau$  agreement with  $\tau$ -risk. Strong, medium and Weak overlap are the tertiles of the overlap measured with NTV eq. 16. G.4 presents results for all metrics in Figure S11 in absolute Kendall's and continuous overlap values in Figure S8.



**Figure 5: Nuisances can be estimated on the same data as outcome models:** Results for the R-risk are similar between the **shared nuisances/candidate set** and the **separated nuisances set** procedures. Figure S10 details results for all metrics.

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

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



**Figure 6: Stacked models are good overall estimators of the nuisances:** Results are shown only for the  $R$ -risk; Figure S12 details every metrics. For Twins, where the true propensity model is linear, stacked and linear estimations of the nuisances performs equivalently, even for a downsampled version ( $N=4,794$ ).

a situation favorable to linear models. In these settings, stacked and linear estimations of the nuisances performs equivalently. Detailed analysis (Figure S13) confirms that using adaptive models –as built by stacking linear models and gradient-boosted trees– suffices to estimate nuisance.

## Discussion and conclusion

**Nuisance models: more gain than pain** Predictive models are increasingly used to reason about treatment effects, for instance in precision medicine to drive individualized decision. Our results highlight that they should be selected, validated, and tuned using different procedures and error measures than those classically used to assess prediction. Rather, selecting the best outcome model according to the  $R$ -risk (eq. Definition 5) leads to more valid causal estimates. Estimating the  $R$ -risk requires a more complex procedure than standard cross-validation used *e.g.* in machine learning: it involves fitting nuisance models necessary for model evaluation. Our results show that these can be learned on the same set of data as the outcome models evaluated. The nuisance models must be well estimated (Figure 6). Our results show that using for nuisance models a flexible stacking-based family of estimator suffices for good model selection. To select propensity score models, we used the Brier score, minimized by the true individual probability. An easy mistake is to use calibration errors popular in machine learning<sup>57,92,51,45</sup> as these select not for the individual posterior probability but for an aggregate error rate<sup>55</sup>.

**More  $R$ -risk to select models driving decisions** Increasingly complex prediction models integrating richer medical data have flourished because their predictions can be easily demonstrated and validated on left-out data. But using them to underpin a decision on whether to treat or not requires more careful validation, using a metric accounting for the putative intervention, the  $R$ -risk. The  $R$ -risk brings a

sizeable benefit to select the most adequate model, even when model development is based on treated and untreated population with little differences, as in RCTs. To facilitate better model selection, we provide Python code<sup>4</sup>. This model-selection procedure puts no constraints on the models used to build predictive models: it opens the door to evaluating a wide range of models, from gradient boosting to convolutional neural, or language models.

## Acknowledgments

We acknowledge fruitful discussions with Bénédicte Colnet.

## Financial disclosure

None reported.

## Conflict of interest

The authors declare no potential conflict of interests.

## References

- [1] Alaa, Ahmed and Schaar, Mihaela Van Der. “Validating Causal Inference Models via Influence Functions”. In: *International Conference on Machine Learning* (May 24, 2019), pp. 191–201.
- [2] Almond, Douglas, Chay, Kenneth Y., and Lee, David S. “The Costs of Low Birth Weight”. In: *The Quarterly Journal of Economics* 120.3 (2005).
- [3] Altman, Douglas G, Vergouwe, Yvonne, Royston, Patrick, and Moons, Karel GM. “Prognosis and prognostic research: validating a prognostic model”. In: *Bmj* 338 (2009).
- [4] Athey, Susan and Imbens, Guido. “Recursive partitioning for heterogeneous causal effects”. In: *Proceedings of the National Academy of Sciences* 113.27 (2016), pp. 7353–7360.
- [5] Athey, Susan, Tibshirani, Julie, and Wager, Stefan. “Generalized random forests”. In: *Annals of Statistics* 47.2 (Apr. 2019), pp. 1148–1178.

---

<sup>4</sup>[https://github.com/soda-inria/causal\\_model\\_selection](https://github.com/soda-inria/causal_model_selection)

- [6] Austin, Peter C. “An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies”. In: *Multivariate Behavioral Research* 3 (May 2011), pp. 399–424.
- [7] Austin, Peter C. and Stuart, Elizabeth 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”. In: *Statistics in Medicine* 34.28 (Dec. 10, 2015), pp. 3661–3679.
- [8] Battocchi, Keith, Dillon, Eleanor, Hei, Maggie, Lewis, Greg, Oka, Paul, Oprescu, Miruna, and Syrgkanis, Vasilis. *EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation*. <https://github.com/microsoft/EconML>.
- [9] Beam, Andrew L and Kohane, Isaac S. “Big data and machine learning in health care”. In: *Jama* 319.13 (2018), pp. 1317–1318.
- [10] Blakely, Tony, Lynch, John, Simons, Koen, Bentley, Rebecca, and Rose, Sherri. “Reflection on modern methods: when worlds collide—prediction, machine learning and causal inference”. In: *International journal of epidemiology* 49.6 (2020), pp. 2058–2064.
- [11] Bouthillier, Xavier, Delaunay, Pierre, Bronzi, Mirko, Trofimov, Assya, Nichyporuk, Brennan, Szeto, Justin, Mohammadi Sepahvand, Nazanin, Raff, Edward, Madan, Kanika, Voleti, Vikram, Ebrahimi Kahou, Samira, Michalski, Vincent, Arbel, Tal, Pal, Chris, Varoquaux, Gael, and Vincent, Pascal. “Accounting for Variance in Machine Learning Benchmarks”. In: *Proceedings of Machine Learning and Systems* 3 (Mar. 15, 2021), pp. 747–769.
- [12] Casucci, Sabrina, Lin, Li, Hewner, Sharon, and Nikolaev, Alexander. “Estimating the causal effects of chronic disease combinations on 30-day hospital readmissions based on observational Medicaid data”. In: *Journal of the American Medical Informatics Association* 25.6 (2018), pp. 670–678.
- [13] Charlson, Mary E., Pompei, Peter, Ales, Kathy L., and MacKenzie, C.Ronald. “A new method of classifying prognostic comorbidity in longitudinal studies: Development and validation”. In: *Journal of Chronic Diseases* 40.5 (Jan. 1987), pp. 373–383.
- [14] Chernozhukov, Victor, Chetverikov, Denis, Demirer, Mert, Duflo, Esther, Hansen, Christian, Newey, Whitney, and Robins, James. “Double/Debiased Machine Learning for Treatment and Structural Parameters”. In: *The Econometrics Journal* (2018), p. 71.



- [15] Curth, Alicia, Svensson, David, and Weatherall, James. “Really Doing Great at Estimating CATE? A Critical Look at ML Benchmarking Practices in Treatment Effect Estimation”. In: *Neurips Process 2021* (2021), p. 14.
- [16] D’Amour, Alexander, Ding, Peng, Feller, Avi, Lei, Lihua, and Sekhon, Jasjeet. “Overlap in observational studies with high-dimensional covariates”. In: *Journal of Econometrics* 221.2 (2021), pp. 644–654.
- [17] Desai, Rishi J, Wang, Shirley V, Vaduganathan, Muthiah, Evers, Thomas, and Schneeweiss, Sebastian. “Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes”. In: *JAMA network open* 3.1 (2020), e1918962–e1918962.
- [18] Dorie, Vincent, Hill, Jennifer, Shalit, Uri, Scott, Marc, and Cervone, Dan. “Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition”. In: *Statistical Science* 34.1 (2019), pp. 43–68.
- [19] Fang, Gang, Annis, Izabela E, Elston-Lafata, Jennifer, and Cykert, Samuel. “Applying machine learning to predict real-world individual treatment effects: insights from a virtual patient cohort”. In: *Journal of the American Medical Informatics Association* 26.10 (2019), pp. 977–988.
- [20] Fontana, Mark Alan, Lyman, Stephen, Sarker, Gourab K, Padgett, Douglas E, and MacLean, Catherine H. “Can machine learning algorithms predict which patients will achieve minimally clinically important differences from total joint arthroplasty?” In: *Clinical orthopaedics and related research* 477.6 (2019), p. 1267.
- [21] Gao, Zijun, Hastie, Trevor, and Tibshirani, Robert. “Assessment of heterogeneous treatment effect estimation accuracy via matching”. In: *Statistics in Medicine* 17 (2021).
- [22] Gretton, Arthur, Borgwardt, Karsten M, Rasch, Malte J, Schölkopf, Bernhard, and Smola, Alexander. “A kernel two-sample test”. In: *The Journal of Machine Learning Research* 13.1 (2012), pp. 723–773.
- [23] Grose, Elysia, Wilson, Samuel, Barkun, Jeffrey, Bertens, Kimberly, Martel, Guillaume, Balaa, Fady, and Khalil, Jad Abou. “Use of Propensity Score Methodology in Contemporary High-Impact Surgical Literature”. In: *Journal of the American College of Surgeons* 230.1 (Jan. 1, 2020), 101–112.e2.

- [24] Gruber, Susan and van der Laan, Mark J. “tmle: An R Package for Targeted Maximum Likelihood Estimation”. In: *Journal of Statistical Software* 51.13 (2012), pp. 1–35.
- [25] Gutierrez, Pierre and Gerardy, Jean-Yves. “Causal Inference and Uplift Modeling A review of the literature”. In: *Proceedings of The 3rd International Conference on Predictive Applications and APIs*. Proceedings of Machine Learning Research 67 (2016), p. 14.
- [26] Hernán, MA and Robins, JM. *Causal Inference: What If*. CRC Boca Raton, FL, 2020.
- [27] Hill, Jennifer L. “Bayesian Nonparametric Modeling for Causal Inference”. In: *Journal of Computational and Graphical Statistics* 1 (Jan. 1, 2011), pp. 217–240.
- [28] Hoogland, Jeroen, IntHout, Joanna, Belias, Michail, Rovers, Maroeska M, Riley, Richard D, E. Harrell Jr, Frank, Moons, Karel GM, Debray, Thomas PA, and Reitsma, Johannes B. “A tutorial on individualized treatment effect prediction from randomized trials with a binary endpoint”. In: *Statistics in medicine* 40.26 (2021), pp. 5961–5981.
- [29] Horng, Steven, Sontag, David A, Halpern, Yoni, Jernite, Yacine, Shapiro, Nathan I, and Nathanson, Larry A. “Creating an automated trigger for sepsis clinical decision support at emergency department triage using machine learning”. In: *PloS one* 12.4 (2017), e0174708.
- [30] Howe, Chanelle J., Cole, Stephen R., Westreich, Daniel J., Greenland, Sander, Napravnik, Sonia, and Eron, Joseph J. “Splines for trend analysis and continuous confounder control”. In: *Epidemiology (Cambridge, Mass.)* 22.6 (Nov. 2011), pp. 874–875.
- [31] Imbens, Guido W. and Rubin, Donald B. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.
- [32] Ionides, Edward L. “Truncated Importance Sampling”. In: *Journal of Computational and Graphical Statistics* 17.2 (June 2008), pp. 295–311.
- [33] Jesson, Andrew, Mindermann, Sören, Shalit, Uri, and Gal, Yarin. “Identifying Causal-Effect Inference Failure with Uncertainty-Aware Models”. In: *Advances in Neural Information Processing Systems* 33 (Oct. 22, 2020), pp. 11637–11649.
- [34] Johansson, Fredrik D, Shalit, Uri, Kallus, Nathan, and Sontag, David. “Generalization bounds and representation learning for estimation of potential outcomes and causal effects”. In: *The Journal of Machine Learning Research* 23.1 (2022), pp. 7489–7538.

- [35] Kendall, M. G. “A new measure of rank correlation”. In: *Biometrika* 30.1-2 (June 1938), pp. 81–93.
- [36] Khojaste-Sarakhsi, M, Haghighi, Seyedhamidreza Shahabi, Ghomi, SMT Fatemi, and Marchiori, Elena. “Deep learning for Alzheimer’s disease diagnosis: A survey”. In: *Artificial Intelligence in Medicine* (2022), p. 102332.
- [37] Künzel, Sören R., Sekhon, Jasjeet S., Bickel, Peter J., and Yu, Bin. “Metalearners for estimating heterogeneous treatment effects using machine learning”. In: *Proceedings of the National Academy of Sciences* 116.10 (Mar. 5, 2019), pp. 4156–4165.
- [38] Laan, Mark J van der, Laan, MJ, and Robins, JM. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.
- [39] Laan, Mark J. van der, Polley, Eric C., and Hubbard, Alan E. “Super Learner”. In: *Statistical Applications in Genetics and Molecular Biology* 6 (Sept. 16, 2007).
- [40] Laan, Mark J. van der and Rose, Sherri. *Targeted Learning*. Springer Series in Statistics. 2011.
- [41] Lamont, Andrea, Lyons, Michael D, Jaki, Thomas, Stuart, Elizabeth, Feaster, Daniel J, Tharmaratnam, Kukatharmini, Oberski, Daniel, Ishwaran, Hemant, Wilson, Dawn K, and Van Horn, M Lee. “Identification of predicted individual treatment effects in randomized clinical trials”. In: *Statistical methods in medical research* 27.1 (2018), pp. 142–157.
- [42] Loiseau, Nicolas, Trichelair, Paul, He, Maxime, Andreux, Mathieu, Zaslavskiy, Mikhail, Wainrib, Gilles, and Blum, Michael G. B. “External control arm analysis: an evaluation of propensity score approaches, G-computation, and doubly debiased machine learning”. In: *BMC Medical Research Methodology* 22 (2022).
- [43] Louizos, Christos, Shalit, Uri, Mooij, Joris, Sontag, David, Zemel, Richard, and Welling, Max. “Causal Effect Inference with Deep Latent-Variable Models”. In: *Advances in neural information processing systems* (Nov. 6, 2017).
- [44] MacDorman, M. F. and Atkinson, J. O. “Infant mortality statistics from the linked birth/infant death data set–1995 period data”. eng. In: *Monthly Vital Statistics Report* 46.6 Suppl 2 (Feb. 1998), pp. 1–22.
- [45] Minderer, Matthias, Djolonga, Josip, Romijnders, Rob, Hubis, Frances, Zhai, Xiaohua, Houlsby, Neil, Tran, Dustin, and Lucic, Mario. “Revisiting the Calibration of Modern Neural Networks”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 15682–15694.

- [46] Mooney, Stephen J and Pejaver, Vikas. “Big data in public health: terminology, machine learning, and privacy”. In: *Annual review of public health* 39 (2018), p. 95.
- [47] Moons, Karel GM, Royston, Patrick, Vergouwe, Yvonne, Grobbee, Diederick E, and Altman, Douglas G. “Prognosis and prognostic research: what, why, and how?” In: *Bmj* 338 (2009).
- [48] Naimi, Ashley I, Mishler, Alan E, and Kennedy, Edward H. “Challenges in Obtaining Valid Causal Effect Estimates with Machine Learning Algorithms”. In: *American Journal of Epidemiology* (2021).
- [49] Naimi, Ashley I and Whitcomb, Brian W. “Defining and Identifying Average Treatment Effects”. In: *American Journal of Epidemiology* (2023).
- [50] Nassif, Ali Bou, Talib, Manar Abu, Nasir, Qassim, Afadar, Yaman, and Elgendy, Omar. “Breast cancer detection using artificial intelligence techniques: A systematic literature review”. In: *Artificial Intelligence in Medicine* (2022), p. 102276.
- [51] Niculescu-Mizil, Alexandru and Caruana, Rich. “Predicting good probabilities with supervised learning”. en. In: *Proceedings of the 22nd international conference on Machine learning - ICML '05*. ACM Press, 2005, pp. 625–632.
- [52] Nie, Xinkun and Wager, Stefan. “Quasi-Oracle Estimation of Heterogeneous Treatment Effects”. In: *Biometrika* 108.2 (Dec. 13, 2017), pp. 299–319.
- [53] Niswander, Kenneth R. and Stroke, United States National Institute of Neurological Diseases and. *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.
- [54] Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, Vanderplas, Jake, Passos, Alexandre, Cournapeau, David, Brucher, Matthieu, Perrot, Matthieu, and Duchesnay, Édouard. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12.85 (2011), pp. 2825–2830.
- [55] Perez-Lebel, Alexandre, Morvan, Marine Le, and Varoquaux, Gaël. “Beyond calibration: estimating the grouping loss of modern neural networks”. In: *arXiv preprint arXiv:2210.16315* (2022).

- [56] Perperoglou, Aris, Sauerbrei, Willi, Abrahamowicz, Michal, and Schmid, Matthias. “A review of spline function procedures in R”. In: *BMC Medical Research Methodology* 19.1 (Mar. 6, 2019), p. 46.
- [57] Platt, John C. and Platt, John C. “Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods”. In: *Advances in Large Margin Classifiers* (1999), pp. 61–74.
- [58] Poldrack, Russell A, Huckins, Grace, and Varoquaux, Gael. “Establishment of best practices for evidence for prediction: a review”. In: *JAMA psychiatry* 77.5 (2020), pp. 534–540.
- [59] Powers, Scott, Qian, Junyang, Jung, Kenneth, Schuler, Alejandro, Shah, Nigam H., Hastie, Trevor, and Tibshirani, Robert. “Some methods for heterogeneous treatment effect estimation in high dimensions”. In: *Statistics in Medicine* 37.11 (2018), pp. 1767–1787.
- [60] Rahimi, Ali and Recht, Benjamin. “Random Features for Large-Scale Kernel Machines”. In: *Advances in Neural Information Processing Systems*. Vol. 20. 2008.
- [61] Rajkomar, Alvin, Dean, Jeffrey, and Kohane, Isaac. “Machine learning in medicine”. In: *New England Journal of Medicine* 380.14 (2019), pp. 1347–1358.
- [62] Robins, James. “A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect”. In: *Mathematical Modelling* 7.9 (Jan. 1, 1986), pp. 1393–1512.
- [63] Robins, James M. and Greenland, Sander. “The role of model selection in causal inference from non experimental data”. In: *American Journal of Epidemiology* 123.3 (Mar. 1986), pp. 392–402.
- [64] Robinson, P. M. “Root-N-Consistent Semiparametric Regression”. In: *Econometrica* 4 (1988), pp. 931–954.
- [65] Rolling, Craig A. and Yang, Yuhong. “Model selection for estimating treatment effects”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.4 (Sept. 2014), pp. 749–769.
- [66] Rosenbaum, Paul R and Rubin, Donald B. “The central role of the propensity score in observational studies for causal effects”. In: *Biometrika* 70 (1983), pp. 41–55.
- [67] Rubin, Donald B. “Causal Inference Using Potential Outcomes”. In: *Journal of the American Statistical Association* 100.469 (Mar. 1, 2005), pp. 322–331.

- [68] Saito, Yuta and Yasui, Shota. “Counterfactual Cross-Validation: Stable Model Selection Procedure for Causal Inference Models”. In: *International Conference on Machine Learning*. PMLR. 2020, pp. 8398–8407.
- [69] Schulam, Peter and Saria, Suchi. “Reliable Decision Support using Counterfactual Models”. In: *Advances in neural information processing systems* 30 (Mar. 30, 2017).
- [70] Schuler, Alejandro, Baiocchi, Michael, Tibshirani, Robert, and Shah, Nigam. “A comparison of methods for model selection when estimating individual treatment effects”. In: *arXiv:1804.05146 [cs, stat]* (June 13, 2018).
- [71] Schuler, Megan S. and Rose, Sherri. “Targeted Maximum Likelihood Estimation for Causal Inference in Observational Studies”. In: *American Journal of Epidemiology* 185.1 (Jan. 1, 2017), pp. 65–73.
- [72] Shalit, Uri, Johansson, Fredrik D, and Sontag, David. “Estimating individual treatment effect: generalization bounds and algorithms”. In: *International Conference on Machine Learning*. PMLR. 2017, pp. 3076–3085.
- [73] Shen, Li, Margolies, Laurie R, Rothstein, Joseph H, Fluder, Eugene, McBride, Russell, and Sieh, Weiva. “Deep learning to improve breast cancer detection on screening mammography”. In: *Scientific reports* 9.1 (2019), p. 12495.
- [74] Shen, Lingjie, Geleijnse, Gijs, and Kaptein, Maurits. “RCTrep: An R Package for the Validation of Estimates of Average Treatment Effects”. In: *Journal of Statistical Software* (2023).
- [75] Shimoni, Yishai, Karavani, Ehud, Ravid, Sivan, Bak, Peter, Ng, Tan Hung, Alford, Sharon Hensley, Meade, Denise, and Goldschmidt, Yaara. “An Evaluation Toolkit to Guide Model Selection and Cohort Definition in Causal Inference”. In: *arXiv preprint arXiv:1906.00442* (2019).
- [76] Shimoni, Yishai, Yanover, Chen, Karavani, Ehud, and Goldschmidt, Yaara. “Benchmarking Framework for Performance-Evaluation of Causal Inference Analysis”. In: *arXiv:1802.05046 [cs, stat]* (Mar. 20, 2018).
- [77] Simon, Gregory E, Johnson, Eric, Lawrence, Jean M, Rossom, Rebecca C, Ahmedani, Brian, Lynch, Frances L, Beck, Arne, Waitzfelder, Beth, Ziebell, Rebecca, Penfold, Robert B, et al. “Predicting suicide attempts and suicide deaths following outpatient visits using electronic health records”. In: *American Journal of Psychiatry* 175.10 (2018), pp. 951–960.

- [78] Snowden, Jonathan M., Rose, Sherri, and Mortimer, Kathleen M. “Implementation of G-computation on a simulated data set: demonstration of a causal inference technique”. eng. In: *American Journal of Epidemiology* 173.7 (Apr. 2011), pp. 731–738.
- [79] Spasic, Irena, Nenadic, Goran, et al. “Clinical text data in machine learning: systematic review”. In: *JMIR medical informatics* 8.3 (2020), e17984.
- [80] Sperrin, Matthew, Jenkins, David, Martin, Glen P, and Peek, Niels. “Explicit causal reasoning is needed to prevent prognostic models being victims of their own success”. In: *Journal of the American Medical Informatics Association* 26.12 (2019), pp. 1675–1676.
- [81] Sriperumbudur, Bharath K., Fukumizu, Kenji, Gretton, Arthur, Schölkopf, Bernhard, and Lanckriet, Gert R. G. “On integral probability metrics,  $\phi$ -divergences and binary classification”. In: *arXiv:0901.2698 [cs, math]* (Oct. 12, 2009).
- [82] Steyerberg, Ewout W. *Clinical prediction models*. Springer, 2019.
- [83] Su, Xiaogang, Peña, Annette T, Liu, Lei, and Levine, Richard A. “Random forests of interaction trees for estimating individualized treatment effects in randomized trials”. In: *Statistics in medicine* 37.17 (2018), pp. 2547–2560.
- [84] Swaminathan, Adith and Joachims, Thorsten. “Counterfactual risk minimization: Learning from logged bandit feedback”. In: *International Conference on Machine Learning*. PMLR. 2015, pp. 814–823.
- [85] VanderWeele, Tyler J. “Principles of confounder selection”. In: *European journal of epidemiology* 34 (2019), pp. 211–219.
- [86] Varoquaux, Gaël and Colliot, Olivier. *Evaluating machine learning models and their diagnostic value*. 2022.
- [87] Wager, Stefan and Athey, Susan. “Estimation and Inference of Heterogeneous Treatment Effects using Random Forests”. In: *Journal of the American Statistical Association* 113.523 (July 3, 2018), pp. 1228–1242.
- [88] Wang, Hanyin, Li, Yikuan, Khan, Seema A, and Luo, Yuan. “Prediction of breast cancer distant recurrence using natural language processing and knowledge-guided convolutional neural network”. In: *Artificial intelligence in medicine* 110 (2020), p. 101977.

- [89] Wendling, T., Jung, K., Callahan, A., Schuler, A., Shah, N. H., and Gallego, B. “Comparing methods for estimation of heterogeneous treatment effects using observational data from health care databases”. In: *Statistics in Medicine* 23 (2018), pp. 3309–3324.
- [90] Wyatt, Jeremy C and Altman, Douglas G. “Commentary: Prognostic models: clinically useful or quickly forgotten?” In: *Bmj* 311.7019 (1995), pp. 1539–1541.
- [91] Yala, Adam, Lehman, Constance, Schuster, Tal, Portnoi, Tally, and Barzilay, Regina. “A deep learning mammography-based model for improved breast cancer risk prediction”. In: *Radiology* 292.1 (2019), pp. 60–66.
- [92] Zadrozny, Bianca and Elkan, Charles. “Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers”. In: (2001), p. 8.
- [93] Zhang, Zhenwei and Sejdić, Ervin. “Radiological images and machine learning: trends, perspectives, and prospects”. In: *Computers in biology and medicine* 108 (2019), pp. 354–370.