Validating Causal Inference Models via Influence Functions

Ahmed M. Alaa [1] **Mihaela van der Schaar** [1 2 3]

Abstract

The problem of estimating causal effects of treatments from observational data falls beyond the realm of supervised learning — because counterfactual data is inaccessible, we can never observe the true causal effects. In the absence of "supervision", *how can we evaluate the performance of causal inference methods?*

In this paper, we use influence functions — the functional derivatives of a loss function — to develop a model validation procedure that estimates the estimation error of causal inference methods. Our procedure utilizes a Taylor-like expansion to approximate the loss function of a method on a given dataset in terms of the influence functions of its loss on a "synthesized", proximal dataset with known causal effects. Under minimal regularity assumptions, we show that our procedure is $\sqrt{n}$-consistent and efficient. Experiments on 77 benchmark datasets show that using our procedure, we can accurately predict the comparative performances of state-of-the-art causal inference methods applied to a given observational study.

1. Introduction

The problem of estimating individualized causal effects of a treatment from observational data is central in many application domains such as healthcare (Foster et al., 2011), computational advertising (Bottou et al., 2013), and social sciences (Xie et al., 2012). In the past few years, numerous machine learning-based models for causal inference were developed, capitalizing on ideas from representation learning (Yao et al., 2018), multi-task learning (Alaa & van der Schaar, 2018) and adversarial training (Yoon et al., 2018). The literature on machine learning-based causal inference is constantly growing, with various related workshops and competitions being held every year (Dorie et al., 2017).

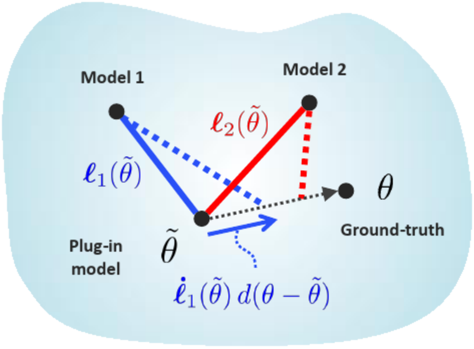

Figure 1. **Pictorial representation of our validation procedure.** To estimate the performances of two competing models (model 1 and model 2), we first estimate their performances with respect to the plug-in distribution ($\ell_1(\tilde{\theta})$ and $\ell_2(\tilde{\theta})$), and then correct for the plug-in bias using the influence functions $\dot{\ell}_1(\tilde{\theta})$ and $\dot{\ell}_2(\tilde{\theta})$.

The fundamental problem of causal inference is that after a subject receives a treatment and displays an outcome, it is impossible to know what the counterfactual outcome would have been had they received an alternative treatment. Since causal effects are determined by both factual and counterfactual outcomes, ground-truth effects can never be measured in an observational study (Stuart et al., 2013). In the absence of "labels" for causal effects, how can we evaluate the performance of causal inference methods?

Addressing this question is an important step for translating advances in (machine learning-based) causal inference into practice. This is because the performance of a given method depends on the dataset at hand, and the comparative performances of different methods can vary wildly across datasets (Dorie et al., 2017). With the vast multitude of methods at their disposal, practitioners need a *data-driven* validation procedure — akin to cross-validation — in order to determine which method to use for a given study. Absent such a procedure, many practitioners would abstain from using machine learning, and instead resort to familiar "white-box" models (e.g., linear regression).

In this paper, we develop a model validation procedure that estimates the performance of causal inference methods applied to a given observational dataset **without the need to access counterfactual data**. To the best of our knowledge,

[1] University of California, Los Angeles, USA [2] University of Cambridge, Cambridge, UK [3] Alan Turing Institute, London, UK. Correspondence to: Ahmed M. Alaa <ahmedmalaa@ucla.edu>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019. Copyright 2019 by the author(s).

ours is the first validation procedure for models of individualized causal effects. Our procedure can be easily extended to other under-explored problems involving unlabeled data, such as semi-supervised learning (Oliver et al., 2018).

In the model validation problem, we are given an observational dataset drawn from an unknown distribution $\mathbb{P}_\theta$ (with a parameter θ), and we want to estimate the model's loss function[1] $\boldsymbol{\ell}(\theta)$ when trained on the dataset at hand. Unlike supervised learning, where we can use cross-validation to estimate $\boldsymbol{\ell}(\theta)$, in causal inference we have no access to the empirical measure of $\boldsymbol{\ell}(\theta)$ because it depends on counterfactual data that we never observe (Shmueli et al., 2010).

Our validation procedure uses influence functions — a key technique in robust statistics and efficiency theory (Hampel et al., 2011; Robins et al., 2008) — to efficiently estimate the loss function of a causal inference model without the need to observe the true causal effects. The key insight behind our validation procedure is that an influence function $\dot{\boldsymbol{\ell}}(\theta)$ of the loss *functional* $\boldsymbol{\ell}(\theta)$ quantifies the "functional derivative" (i.e., Gâteaux derivative) of $\boldsymbol{\ell}(\theta)$ with respect to the data distribution $\mathbb{P}_\theta$ (van der Vaart, 2014). Thus, if we know the model's loss under some *known* distribution $\mathbb{P}_{\tilde{\theta}}$ that is close enough to the true distribution $\mathbb{P}_\theta$, then we can estimate $\boldsymbol{\ell}(\theta)$ via a Taylor expansion as follows:

$$\boldsymbol{\ell}(\theta) \;\approx\; \underset{\substack{\downarrow \\ \textbf{Plug-in estimate}}}{\boxed{\boldsymbol{\ell}(\tilde{\theta})}} \;+\; \underset{\substack{\downarrow \\ \textbf{Plug-in bias}}}{\boxed{\dot{\boldsymbol{\ell}}(\tilde{\theta})\, d(\theta - \tilde{\theta})}}$$

Our two-step validation procedure is succinctly described by the equation above. In the first step, we use the observed data to synthesize a "plug-in" distribution $\mathbb{P}_{\tilde{\theta}}$, under which we can calculate the model's loss $\boldsymbol{\ell}(\tilde{\theta})$. In the second step, we calculate the influence function $\dot{\boldsymbol{\ell}}(\tilde{\theta})$ to correct for the plug-in bias resulting from evaluating the model's loss under $\tilde{\theta}$ instead of θ. (A pictorial depiction of our procedure is provided in Figure 1, where we illustrate its use in validating two competing models[2] for the sake of model selection.) In Section 3, we show that under minimal regularity conditions, our validation approach is consistent, achieves the optimal parametric rate $O_{\mathbb{P}}(n^{-\frac{1}{2}})$, and is efficient (minimizes the variance of its estimates).

To demonstrate the practical significance of our model validation procedure, we collected all causal inference methods published at ICML, NeurIPS and ICLR between 2016 and 2018, and used our procedure to predict their comparative performances on 77 benchmark datasets from a recent causal inference competition (Dorie et al., 2017). We show that using our procedure, practitioners can accurately predict the comparative performances of state-of-the-art methods applied to a given observational study. Thus, epidemiologists and applied statistician can use our procedure to select the right model for the observational study at hand.

Related Work

In the past few years, there has been a notable growth in research developing machine learning methods for estimating individualized causal effects (Zhao et al., 2019; Subbaswamy & Saria, 2018; Johansson et al., 2018; Atan et al., 2018; Yao et al., 2018; Chernozhukov et al., 2018; Ray & van der Vaart, 2018; Künzel et al., 2017; Li & Fu, 2017; Hahn et al., 2017; Powers et al., 2018; Kallus, 2017; Johansson et al., 2016; Shalit et al., 2017). The modeling approaches used in those works were vastly diverse, ranging from Gaussian processes (e.g., (Alaa & van der Schaar, 2017)), to causal random forests (e.g., (Wager & Athey, 2017)) to generative adversarial networks (e.g., GAN-ITE (Yoon et al., 2018)). We present a detailed survey of existing methods in the Supplementary material.

Researchers developing new methods for causal inference validate their models using synthetic data-generating distributions that encode pre-specified causal effects — e.g., (Hill, 2011; Wager & Athey, 2017; Powers et al., 2018). However, such synthetic distributions bear very little resemblance to real-world data, and hence are not informative of what methods would actually work best on a given real-world observational study (Setoguchi et al., 2008). Because no single model will be superior on all observational studies (Dorie et al., 2017), model selection must be guided by a data-driven validation procedure.

While the literature is rich with causal inference models, it falls short of rigorous methods for validating those models on real-world data. Applied researchers currently rely on simple heuristics to predict a model's performance on a given dataset (Schuler et al., 2017; Rolling & Yang, 2014; Van der Laan et al., 2003), but such heuristics do not provide any theoretical guarantees, and can fail badly in certain scenarios (Schuler et al., 2018). In Section 5, we conduct empirical comparisons between our procedure and various commonly-used heuristics.

Despite their popularity in statistics, influence functions are seldom used in machine learning. Recently in (Koh & Liang, 2017), influence functions were used for interpreting black-box models by tracing the impact of data points on a model's predictions. Our usage of influence functions differs from (Koh & Liang, 2017) in that we use them to construct efficient estimators of a model's loss and not to explain the inner workings of a learning algorithm. In that sense, our work is more connected to the literature on plug-in estimation and nonparametric efficiency theory (Goldstein & Messer, 1992; Robins et al., 2008; 2017; van der Vaart, 2014).

[1] We use simplified notation in this Section for ease of exposition. Precise notational definitions are provided in Section 2.

[2] In Figure 1, we do not show $\dot{\boldsymbol{\ell}}_2(\tilde{\theta})$ to avoid clutter.

2. Problem Setup

2.1. Causal Inference from Observational Data

We consider the standard *potential outcomes* framework for modeling causal effects in observational and experimental studies (Rubin, 1974; 2005). In this framework, a "subject" is associated with a feature $X \in \mathcal{X}$, a treatment assignment indicator $W \in \{0, 1\}$, and an outcome $Y \in \mathbb{R}$. The outcome variable Y takes on the value of either of the two "potential outcomes" $Y^{(0)}$ and $Y^{(1)}$, where $Y = Y^{(1)}$ if the subject received the treatment ($W = 1$), and $Y = Y^{(0)}$ otherwise, i.e., $Y = W\,Y^{(1)} + (1 - W)\,Y^{(0)}$. The causal effect of the treatment on the subject is thus given by $Y^{(1)} - Y^{(0)}$.

■ **Observational data.** In a typical observational study, we are given n samples of the tuple $Z = (X, W, Y)$ drawn from a probability distribution with a parameter θ, i.e.,

$$Z_1, \ldots, Z_n \sim \mathbb{P}_\theta\,, \tag{1}$$

where $\mathbb{P}_\theta$ belongs to the family $\mathcal{P} = \{\mathbb{P}_{\theta'} : \theta' \in \Theta\}$, and Θ is the parameter space. We break down the parameter θ into a collection of *nuisance* parameters $\theta = \{\mu_0, \mu_1, \pi, \eta\}$, where μ_0 and μ_1 are the conditional potential outcomes, i.e.,

$$\mu_w(x) = \mathbb{E}_\theta\big[\,Y^{(w)}\,|\,X = x\,\big],\; w \in \{0, 1\}, \tag{2}$$

and π is the treatment assignment mechanism, i.e.

$$\pi(x) = \mathbb{P}_\theta(W = 1\,|\,X = x\,), \tag{3}$$

whereas $\eta(x) = \mathbb{P}_\theta(X = x)$. To ensure the generality of our analysis, we assume that $\mathcal{P}$ is a *nonparametric* family of distributions. That is, Θ is an infinite-dimensional parameter space, with the nuisance parameters $\{\mu_0, \mu_1, \pi, \eta\}$ being specified only through mild smoothness conditions.

■ **The causal inference task.** The goal of causal inference is to use the samples $\{Z_i\}_{i=1}^n$ in order to infer the causal effect of the treatment on individual subjects based on their features, i.e., the estimand is a function $T : \mathcal{X} \to \mathbb{R}$, where

$$T(x) = \mathbb{E}_\theta\big[\,Y^{(1)} - Y^{(0)}\,|\,X = x\,\big]. \tag{4}$$

The function $T(x)$ in (4) is commonly known as the conditional average treatment effect (CATE)[3]. Its importance resides in the fact that it can guide *individualized* decision-making policies (e.g., patient-specific treatment plans or personalized advertising policies (Bottou et al., 2013)). For this reason, the CATE function is the estimand of interest for almost all modern machine learning-based causal inference methods (e.g., (Alaa & van der Schaar, 2018; Wager & Athey, 2017; Yoon et al., 2018; Yao et al., 2018)).

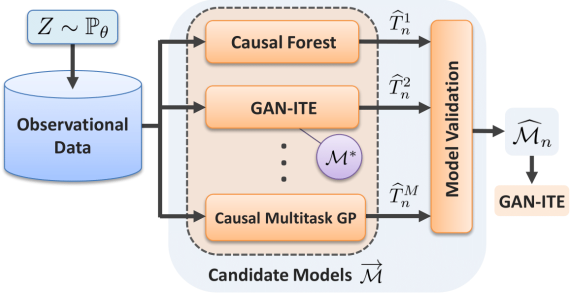

Figure 2. Schematic of the causal inference model validation and selection procedure. In this example, $\mathcal{M}^*$ = GAN-ITE, hence the procedure succeeds if $\widehat{\mathcal{M}}_n$ = GAN-ITE.

■ **Accuracy of causal inference.** A causal inference model $\mathcal{M}$ maps a dataset $\{Z_i\}_{i=1}^n$ to an estimate $\widehat{T}(.)$ of the CATE. The accuracy of a model is typically characterized by the squared-L^2 loss incurred by its estimate, i.e.,

$$\ell_\theta(\widehat{T}) \triangleq \big\|\,\widehat{T}(X) - T(X)\,\big\|_\theta^2, \tag{5}$$

where $\|f(X)\|_\theta^2 = \mathbb{E}_\theta[f^2(X)]$. The performance evaluation metric in (5) was dubbed the *precision of estimating heterogeneous effects* (PEHE) in (Hill, 2011) — it quantifies the ability of a model to capture the heterogeneity of the causal effects of a treatment among individuals in a population.

2.2. Model Validation & Selection

We now consider a set of candidate causal inference models $\overrightarrow{\mathcal{M}} = \{\mathcal{M}_1, \ldots, \mathcal{M}_M\}$. These may include, for example, different machine learning methods (e.g., Causal Gaussian processes, GAN-ITE, causal forests, etc.), different hyperparameter settings of one method, etc. Our goal is to select the best model $\mathcal{M}^* \in \overrightarrow{\mathcal{M}}$ that incurs the minimum PEHE for a given dataset. A schematic depiction of our model selection framework is provided in Figure 2.

■ **Beyond cross-validation.** Evidently, reliable model selection requires a model validation procedure that estimates the PEHE accuracy of each model in $\overrightarrow{\mathcal{M}}$. Unlike standard supervised learning in which all data points are definitely "labeled", in the causal inference setting we do not have access to the ground-truth causal effect $Y^{(1)} - Y^{(0)}$. This is because in an observational dataset, we can only observe the factual outcome $Y^{(W)}$, but not the counterfactual $Y^{(1\,-\,W)}$. This renders the empirical measure of PEHE, i.e., $\frac{1}{n}\sum_{i=1}^n(\widehat{T}(X_i) - (Y_i^{(1)} - Y_i^{(0)}))^2$, incalculable from the samples $\{Z_i = (X_i, W_i, Y_i)\}_{i=1}^n$, and hence standard cross-validation techniques cannot be used to evaluate the performance of a given causal inference model[4].

[3] To ensure the identification of the CATE, we assume that $\mathbb{P}_\theta$ satisfies the standard "unconfoundedness" and "overlap" conditions in (Pearl, 2009; Rubin, 2005).

[4] In Appendix B, we analyze a number of naïve alternatives to cross-validation that were used in previous works to tune the hyperparameter of causal inference models (Shalit et al., 2017; Shimodaira, 2000), etc.). We show that all such alternatives provide either inconsistent or inefficient estimates of the PEHE.

3. Model Validation via Influence Functions

How can we test the PEHE performance of a causal inference model without observing a test label $Y^{(1)} - Y^{(0)}$?

To answer this question, we develop a consistent and efficient validation procedure that estimates the PEHE of any causal inference model via a statistic that **does not depend on the counterfactual data $\boldsymbol{Y^{(1-W)}}$**. Using this procedure, practitioners can evaluate, compare and select causal inference models as envisioned in Section 2.2.

Our validation procedure adopts a *plug-in* estimation principle (Wright et al., 2011), whereby the true (unobserved) causal effect T is replaced with an estimate $\widetilde{T}$. The key idea of our procedure is that — since PEHE is a *functional* of distributions spanned by Θ — if we know a model's PEHE under a proximal plug-in distribution $\mathbb{P}_{\tilde{\theta}} \approx \mathbb{P}_{\theta}$, then we can approximate its true PEHE under $\mathbb{P}_{\theta}$ using a (generalized) Taylor expansion. In such an expansion, the *influence functions* of the PEHE functional are analogous to derivatives of a function in standard calculus.

A high-level description of our procedure is given below.

Input: Observational data, a model $\mathcal{M}$.

1. **Step 1: Plug-in estimation**
   - Fit a *plug-in* model $\tilde{\theta} = \{\tilde{\mu}_0, \tilde{\mu}_1, \tilde{\pi}, \tilde{\eta}\}$.
   - Compute a plug-in estimate $\boldsymbol{\ell}_{\tilde{\theta}}$ of the PEHE.
2. **Step 2: Unplugged validation**
   - Use the influence functions of $\boldsymbol{\ell}_{\tilde{\theta}}$ to predict $\boldsymbol{\ell}_{\theta}$.

Output: An estimate of the PEHE for model $\mathcal{M}$.

In what follows, we provide a detailed explanation of the two-step procedure above.

3.1. Step 1: Plug-in Estimation

In Step 1, we obtain an initial guess of a model's PEHE by evaluating the PEHE functional at an estimated parameter $\tilde{\theta}$ instead of the true parameter θ, i.e.,

$$\boldsymbol{\ell}_{\tilde{\theta}}(\widehat{T}) = \|\, \widehat{T}(X) - \widetilde{T}(X)\, \|^2_{\tilde{\theta}}, \qquad (6)$$

where $\widehat{T}$ is the CATE estimate of the model $\mathcal{M}$ being validated, $\tilde{\theta} = \{\tilde{\mu}_0, \tilde{\mu}_1, \tilde{\pi}, \tilde{\eta}\}$ is a *plug-in model* that is obtained from the observational data, and $\widetilde{T}(x) = \tilde{\mu}_1(x) - \tilde{\mu}_0(x)$.

The plug-in model $\tilde{\theta} = \{\tilde{\mu}_0, \tilde{\mu}_1, \tilde{\pi}, \tilde{\eta}\}$ is estimated from the observational data $\{Z_i\}_{i=1}^n$ as follows:

- $\tilde{\mu}_w$, $w \in \{0, 1\}$, is obtained by fitting a supervised regression model to the sub-dataset $\{(X_i, Y_i) \,|\, W_i = w\}$.
- $\tilde{\pi}$ is obtained using a supervised classification model fit to the sub-dataset $\{(X_i, W_i)\}_{i=1}^n$.

The feature distribution component of $\tilde{\theta}$, $\tilde{\eta}(x)$, can be obtained by estimating the density of X using the feature samples $\{X_i\}_{i=1}^n$. (We defer the implementation details of the plug-in model $\tilde{\theta}$ till Section 5.) Once we have obtained $\tilde{\theta}$, the plug-in PEHE estimate in (6) can be easily evaluated.

The plug-in approach in (6) solves the problem of the inaccessibility of the label $Y^{(1)} - Y^{(0)}$ by "synthesizing" such label through the plug-in model $\tilde{\theta}$, and testing a model's inferences against the synthesized CATE function $\widetilde{T}$. This idea is the basis for recent model selection schemes, such as Synth-Validation (Schuler et al., 2017) and Plasmode simulations (Franklin et al., 2014), which propose similar plug-in approaches for validating causal inference models.

■ **Plug-in estimation is not sufficient.** The plug-in estimate in (6) exhibits a model-dependent plug-in bias $\boldsymbol{\ell}_{\theta} - \boldsymbol{\ell}_{\tilde{\theta}}$ that makes it of little use for model selection. This is because $\boldsymbol{\ell}_{\tilde{\theta}}(\widehat{T})$ measures how well $\widehat{T}$ approximates the synthesized causal effect $\widetilde{T}$ and not the true effect T. Thus, comparing plug-in PEHE estimates of different models can reveal their true comparative performances only if the plug-in bias is small[5], i.e., $\|\widetilde{T} - T\|^2_{\theta} \approx 0$. However, if $\|\widetilde{T} - T\|^2_{\theta}$ is large, then plug-in PEHEs tell us nothing about how different models compare on the true distribution $\mathbb{P}_{\theta}$.

3.2. Step 2: Unplugged Validation

How can we get the plug-in bias "unplugged"? We begin by noting that the plug-in PEHE and the true PEHE are two evaluations of the same functional at $\tilde{\theta}$ and θ, respectively. Therefore, we can write $\boldsymbol{\ell}_{\theta}$ in terms of $\boldsymbol{\ell}_{\tilde{\theta}}$ via a *von Mises* expansion as follows (Fernholz, 2012):

$$\boldsymbol{\ell}_{\theta}(\widehat{T}) = \boldsymbol{\ell}_{\tilde{\theta}}(\widehat{T}) + \sum_{k=1}^{\infty} \int \frac{\dot{\boldsymbol{\ell}}^{(k)}_{\tilde{\theta}}(\boldsymbol{z}; \widehat{T})}{k!}\, d(\mathbb{P}_{\theta} - \mathbb{P}_{\tilde{\theta}})^{\otimes k}, \quad (7)$$

where $\dot{\boldsymbol{\ell}}^{(k)}_{\theta}(\boldsymbol{z}; \widehat{T}) = \dot{\boldsymbol{\ell}}^{(k)}_{\theta}(z_1, \ldots, z_k; \widehat{T})$ is a k^{th} order influence function of the PEHE functional at θ (indexed by $\widehat{T}$), with z being a realization of the variable Z in (1), and $\mathbb{P}^{\otimes k}_{\theta}$ is the k-fold product measure of $\mathbb{P}_{\theta}$.

■ **Influence functions.** The von Mises expansion generalizes Taylor expansion to functionals — it recovers the PEHE at θ based solely on its (higher order) influence functions at $\tilde{\theta}$. In this sense, the influence functions of functionals are analogous to the derivatives of (analytic) functions. Influence functions may not be unique: any set of unbiased k-input functions — i.e., $\mathbb{E}_{\theta}[\dot{\boldsymbol{\ell}}^{(k)}_{\theta}(\boldsymbol{Z}; \widehat{T})] = 0$ — that satisfy (7) are valid influence functions. We discuss how to calculate the influence functions of $\boldsymbol{\ell}_{\tilde{\theta}}$ in Section 4.

[5] Paradoxically enough, if $\widetilde{T}$ is a perfect estimate of T (i.e., $\|\widetilde{T} - T\|^2_{\theta} = 0$), then the model selection task itself becomes obsolete, because the plug-in model would already be better than any model being evaluated. With the plug-in approach, however, we can never know how accurate $\widetilde{T}$ is, since $\boldsymbol{\ell}_{\tilde{\theta}}(\widetilde{T}) = 0$ by definition.

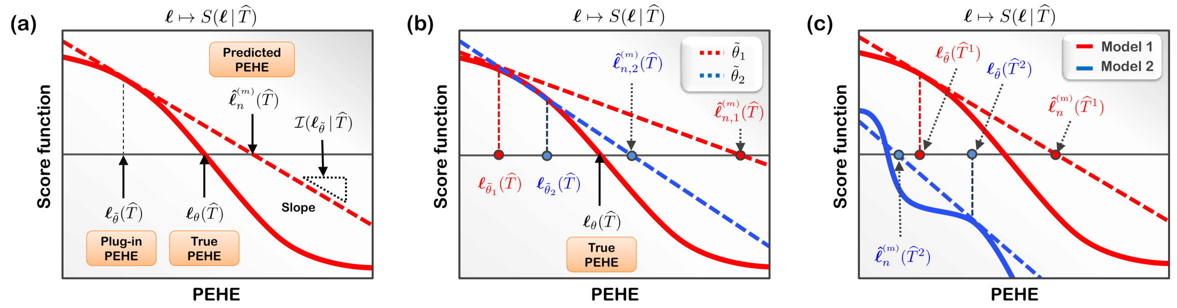
(a)
$\ell \mapsto S(\ell \,|\, \widehat{T})$
Predicted PEHE
$\tilde{\ell}^{(m)}_n(\widehat{T})$
$\mathcal{I}(\ell_{\tilde{\theta}} \,|\, \widehat{T})$
Slope
$\ell_{\tilde{\theta}}(\widehat{T})$
$\ell_{\theta}(\widehat{T})$
Plug-in PEHE
True PEHE
Score function
PEHE
(b)
$\ell \mapsto S(\ell \,|\, \widehat{T})$
$\tilde{\theta}_1$
$\tilde{\theta}_2$
$\hat{\ell}^{(m)}_{n,2}(\widehat{T})$
$\hat{\ell}^{(m)}_{n,1}(\widehat{T})$
$\ell_{\tilde{\theta}_1}(\widehat{T})$
$\ell_{\tilde{\theta}_2}(\widehat{T})$
$\ell_{\theta}(\widehat{T})$
True PEHE
Score function
PEHE
(c)
$\ell \mapsto S(\ell \,|\, \widehat{T})$
Model 1
Model 2
$\ell_{\tilde{\theta}}(\widehat{T}^1)$
$\ell_{\tilde{\theta}}(\widehat{T}^2)$
$\hat{\ell}^{(m)}_n(\widehat{T}^1)$
$\hat{\ell}^{(m)}_n(\widehat{T}^2)$
Score function
PEHE

Figure 3. **Validating causal inference models via influence functions.** Panels (a)-(c) depict exemplary MLE estimating equations for the PEHE as explained in Section 3.3. The x-axis corresponds to PEHE values ($\boldsymbol{\ell}$), and the y-axis corresponds to the score function $S(\boldsymbol{\ell} \,|\, \widehat{T})$. The true PEHE $\boldsymbol{\ell}_\theta(\widehat{T})$ solves the estimating equation $S(\boldsymbol{\ell} \,|\, \widehat{T}) = 0$. Solid lines (—) correspond to $S(\boldsymbol{\ell} \,|\, \widehat{T})$, whereas dashed lines (- - - - -) depict the derivative of the score at the plug-in PEHE. **(a)** The unplugged validation step is analogous to the first iteration of Fisher scoring via Newton-Raphson method. The predicted PEHE is obtained by correcting for the plug-in bias, which is inversely proportional to the Fisher information metric $\mathcal{I}(\boldsymbol{\ell} \,|\, \widehat{T})$. **(b)** Comparison between two plug-in estimates $\tilde{\theta}_1$ and $\tilde{\theta}_2$ for a score function $S(\boldsymbol{\ell} \,|\, \widehat{T})$ (—). The better plug-in estimate conveys more (Fisher) information on the true PEHE, i.e., slope of (- - - - -) is steeper than that of (- - - - -), and hence it provides a better PEHE prediction. **(c)** Selecting between two models $\widehat{T}^1$ and $\widehat{T}^2$ with score functions $S(\boldsymbol{\ell} \,|\, \widehat{T}^1)$ and $S_\theta(\boldsymbol{\ell} \,|\, \widehat{T}^2)$, respectively. While $\widehat{T}^1$ has a smaller plug-in PEHE than $\widehat{T}^2$, predicted PEHEs flip after correcting for plug-in bias.

An influence function $\dot{\boldsymbol{\ell}}^{(k)}_{\tilde{\theta}}(z_1, \ldots, z_k; \widehat{T})$ can be interpreted as a "measure of the dependence of $\boldsymbol{\ell}_{\tilde{\theta}}$ on the value of k data points in the observational data", i.e., its value reflects the sensitivity of the plug-in PEHE estimate to perturbations in the data. Marginalizing out the data $(z_1, \ldots, z_k)$ with respect to $d(\mathbb{P}_\theta - \mathbb{P}_{\tilde{\theta}})$ results in a functional derivative of $\boldsymbol{\ell}_{\tilde{\theta}}$ in the "direction" $(\mathbb{P}_\theta - \mathbb{P}_{\tilde{\theta}})$ (Robins et al., 2017).

The expansion in (7) represents the plug-in bias $\boldsymbol{\ell}_\theta - \boldsymbol{\ell}_{\tilde{\theta}}$ in terms of functional derivatives of $\boldsymbol{\ell}_{\tilde{\theta}}$. To see how the bias is captured, consider the first-order von Mises expansion, i.e.,

$$\boldsymbol{\ell}_\theta(\widehat{T}) \approx \boldsymbol{\ell}_{\tilde{\theta}}(\widehat{T}) + \int \dot{\boldsymbol{\ell}}^{(1)}_{\tilde{\theta}}(z; \widehat{T})\, d(\mathbb{P}_\theta - \mathbb{P}_{\tilde{\theta}}). \qquad (8)$$

Thus, the plug-in bias will be large if the functional derivative of $\boldsymbol{\ell}_{\tilde{\theta}}$ is large, i.e., the PEHE estimate is sensitive to changes in the plug-in model $\tilde{\theta}$. This derivative will be large if many data points have large influence, and for each such data point, the plug-in distribution is not a good representative of the true distribution, i.e., $d(\mathbb{P}_\theta - \mathbb{P}_{\tilde{\theta}})$ is large.

■ **Dispensing with the counterfactuals.** Note that the expansion in (7) quantifies the plug-in bias in terms of fixed functions of "factual" observations $Z = (X, W, Y^{(W)})$ only. Thus, the true PEHE can be estimated without knowledge of the counterfactual outcome $Y^{(1-W)}$ by calculating the sample average of the first m terms of (7) as follows:

$$\hat{\boldsymbol{\ell}}^{(m)}_n(\widehat{T}) = \boldsymbol{\ell}_{\tilde{\theta}}(\widehat{T}) + \sum_{k=1}^{m} \frac{1}{k!}\, \mathbb{U}_n \left[\dot{\boldsymbol{\ell}}^{(k)}_{\tilde{\theta}}(\boldsymbol{Z}; \widehat{T}) \right], \qquad (9)$$

where $\mathbb{U}_n$ is the empirical U-statistic, i.e., the sample average of a multi-input function (see Appendix B). (9) follows directly from (7) by capitalizing on the unbiasedness of influence functions, i.e., $\mathbb{E}_{\tilde{\theta}}[\dot{\boldsymbol{\ell}}^{(k)}_{\tilde{\theta}}(\boldsymbol{Z}; \widehat{T})] = 0$, $\forall k$.

3.3. Relation to Maximum Likelihood Estimation

In Section 3.2, we used functional calculus to construct a mathematical approximation of a model's PEHE that does not depend on counterfactual data. But is this approximation also a *statistically efficient* PEHE estimate?

Recall that in (parametric) maximum likelihood estimation (MLE), a parameter estimate θ^* is obtained by solving the estimating equation $S(\theta) = 0$, where $S(\theta)$ is the *score* function — i.e., the derivative of the log-likelihood function. For estimating equations that cannot be solved analytically, the classical *Fisher scoring* procedure (Longford, 1987) is used to obtain a numerical solution for the MLE.

Our two-step validation procedure[6] is equivalent to finding the MLE of a model's PEHE using the classical Fisher scoring procedure. To illustrate this equivalence, we capture the structural resemblance between the two procedures in Figure 3 as well as the tabulated comparison below.

Estimating equation	Fisher scoring
(Parametric MLE) $S(\theta^*) = 0$	$\hat{\theta} \approx \theta_0 + \boxed{\mathcal{I}^{-1}(\theta_0)\, S(\theta_0)}$
(Our procedure) $S(\boldsymbol{\ell}^* \,\|\, \widehat{T}) = 0$	$\boldsymbol{\ell}_\theta(\widehat{T}) \approx \boldsymbol{\ell}_{\tilde{\theta}}(\widehat{T}) + \boxed{\mathbb{E}_\theta[\dot{\boldsymbol{\ell}}^{(1)}_{\tilde{\theta}}(\boldsymbol{z}; \widehat{T})]}$

Fisher scoring implements the Newton-Raphson numerical method to solve $S(\theta) = 0$. It utilizes the Taylor approximation of $S(\theta)$ around an initial θ_0 to formulate an iterative equation $\hat{\theta}_{k+1} = \theta^k + \mathcal{I}^{-1}(\theta_k)\, S(\theta_k)$ — where $\mathcal{I}(\theta)$ is the *Fisher information* — that eventually converges to θ^*.

[6] Here we consider a first-order von Mises approximation.

From the tabulated comparison above, we can see that our procedure is analogous to the first Newton-Raphson iteration of Fisher scoring. That is, the plug-in estimation step is similar to finding an initial estimate θ_0, and the unplugged validation step is similar to updating the initial estimate.

This analogy suggests that our procedure is statistically sound. Similar to cross-validation in supervised learning (Dudoit & van der Laan, 2005), our procedure is a de facto MLE algorithm that computes the "most likely PEHE of a model given observational data". As shown in Figure 3-(a), it does so by searching for the root of the score $S(\boldsymbol{\ell}\,|\,\widehat{T})$ via a one-shot Newton-Raphson procedure.

The juxtaposition of our procedure and Fisher scoring — in the tabulated comparison above — suggests an operational definition for Fisher information $\mathcal{I}(\boldsymbol{\ell}\,|\,\widehat{T})$ as the ratio between the score function and influence function. (This relation also holds in parametric models (Basu et al., 1998).) The expression of the plug-in bias in terms of the Fisher metric provides an information-geometric interpretation of our validation procedure. That is, the Fisher information content of the plug-in model $\tilde{\theta}$ determines how much the final PEHE estimate will deviate from the initial plug-in estimate (see Figures 3-(b) and 3-(c) for depictions).

3.4. Consistency and Efficiency

In the following Theorem[7], we establish the conditions under which our validation procedure is statistically efficient.

Theorem 1. *Let μ_0, μ_1, and π be bounded Hölder functions with Hölder exponents α_0, α_1 and β, respectively, and $X \in [0,1]^d$. If (i) $\widehat{T}$ and $\tilde{\theta}$ are fit using a separate sample than that used to compute $\hat{\boldsymbol{\ell}}_n^{(m)}(\widehat{T})$, and (ii) $\widetilde{T}$ is a minimax optimal estimate of T, then we have that:*

$$\hat{\boldsymbol{\ell}}_n^{(m)}(\widehat{T}) - \boldsymbol{\ell}_\theta(\widehat{T}) = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} \vee n^{\frac{-(\alpha_0 \wedge \alpha_1)(m+1)}{2(\alpha_0 \wedge \alpha_1)+d}}\right).$$

If $m \geq \lceil \frac{d}{2(\alpha_0 \wedge \alpha_1)} \rceil$, then the following is satisfied:

$$\textbf{(Consistency)} \quad \sqrt{n}\,(\hat{\boldsymbol{\ell}}_n^{(m)}(\widehat{T}) - \boldsymbol{\ell}_\theta(\widehat{T})) \xrightarrow{d} \mathcal{N}(0,\sigma^2),$$

$$\textbf{(Efficiency)} \quad \mathrm{Var}[\hat{\boldsymbol{\ell}}_n^{(m)}(\widehat{T})] \leq \mathrm{Var}[\hat{\boldsymbol{\ell}}'(\widehat{T})],$$

for some constant $\sigma > 0$, and any estimator $\hat{\boldsymbol{\ell}}'(\widehat{T})$. □

This result gives a cut-off value on the minimum number of influence terms m needed for the PEHE estimator $\hat{\boldsymbol{\ell}}_n^{(m)}(\widehat{T})$ to be statistically efficient. This cut-off value depends on the dimensionality and smoothness of the CATE function.

Theorem 1 also says that the plug-in model $\tilde{\theta}$ needs to be good enough for our procedure to work, i.e., $\widetilde{T}$ must be a minimax optimal approximation of T. This is a viable requirement: it is satisfied by models such as Gaussian processes and regression trees (Alaa & van der Schaar, 2018). Finally, Theorem 1 also says that our procedure yields the minimum variance estimator of a model's PEHE. This can be understood in the light of the analogy with MLE (Section 3.3): since influence functions are proportional to Fisher information, the PEHE estimate in (9) satisfies the Cramér-Rao lower bound on estimation variance.

4. Calculating Influence Functions

Recall that influence functions operationalize the derivatives of $\boldsymbol{\ell}_\theta(.)$ with respect to distributions induced by θ. But since $\mathbb{P}_\theta$ is nonparametric — i.e., θ is infinite-dimensional — how can we compute such derivatives?

A common approach for computing the influence functions of a functional of a nonparametric family $\mathcal{P}$ is to define a smooth parametric submodel of $\mathcal{P}$, and then differentiate the functional with respect to the submodel's (scalar) parameter (van der Vaart, 2014; Kennedy, 2018). A parametric submodel $\mathcal{P}_\varepsilon = \{\mathbb{P}_\varepsilon : \varepsilon \in \mathbb{R}\} \subset \mathcal{P}$ is a subset of models in $\mathcal{P}$ that coincides with $\mathbb{P}_\theta$ at $\varepsilon = 0$. In this paper, we choose to work with the following parametric submodel: $d\mathbb{P}_\varepsilon(z) = (1 + \varepsilon h(z))\, d\mathbb{P}_\theta(z)$, for a bounded function $h(z)$.

Given the submodel $\mathbb{P}_\varepsilon$, it can be shown (by manipulating the von Mises series in (7) — see Appendix G) that the first order influence function satisfies the following condition:

$$\left.\frac{\partial\, \boldsymbol{\ell}_\varepsilon(\widehat{T})}{\partial \varepsilon}\right|_{\varepsilon=0} = \mathbb{E}_\theta[\, \dot{\boldsymbol{\ell}}_\theta^{(1)}(z;\widehat{T}) \cdot S_\varepsilon(z)|_{\varepsilon=0}\,], \qquad (10)$$

where $S_\varepsilon(z) = \partial \log(d\mathbb{P}_\varepsilon(z))/\partial\varepsilon$ is the score function of the parametric submodel, and $\boldsymbol{\ell}_\varepsilon$ is the PEHE functional evaluated at $\mathbb{P}_\varepsilon$. In the next Theorem, we use (10) to derive the closed-form expression for $\dot{\boldsymbol{\ell}}_\theta^{(1)}(z;\widehat{T})$.

Theorem 2. *The first order influence function of the PEHE functional $\boldsymbol{\ell}_\theta(\widehat{T})$ is unique, and is given by:*

$$\begin{aligned}\dot{\boldsymbol{\ell}}_\theta^{(1)}(Z;\widehat{T}) = (1-B)\,T^2(X) + B\,Y\,(T(X) - \widehat{T}(X)) - \\ A\,(T(X) - \widehat{T}(X))^2 + \widehat{T}^2(X) - \boldsymbol{\ell}_\theta(\widehat{T}),\end{aligned}$$

where $A = (W - \pi(X))$, and $B = 2W\,(W - \pi(X)) \cdot C^{-1}$ for $C = \pi(X)(1-\pi(X))$. □

This result implies that the influence functions of $\boldsymbol{\ell}_\theta(\widehat{T})$ do not depend on $\eta(x)$. Thus, the plug-in model $\tilde{\theta}$ does not need to be generative. This is a great relief since estimating (high-dimensional) feature distributions can be daunting.

■ **Influence functions of influence functions.** Unfortunately, higher order influence functions of PEHE are intractable, hence closed-form expressions akin to Theorem 2 are not feasible. In Appendix D, we propose a recursive finite difference algorithm to approximate higher order influence using the fact that influence functions are derivatives, hence higher order influence is obtained by calculating first order influence of lower order influence functions.

[7] All proofs are provided in the Appendix.

5. Automated Causal Inference: A Case Study

As envisioned in Figure 2, practitioners can use our validation procedure to select the best causal inference method for a given dataset. Unlike pervasive "expert-driven" modeling practices (Rubin, 2010), this *automated* and data-driven approach to model selection enables confident deployment of (black-box) machine learning-based methods, and safeguards against naïve modeling choices.

In this Section, we demonstrate the practical significance of influence function-based validation by assessing its utility in model selection. In particular, we assemble a pool of models — comprising all methods published recently in ICML, NeurIPS and ICLR — and use our validation procedure to predict the best performing model on 77 benchmark datasets from a recent causal inference competition.

5.1. Experimental Setup

■ **Influence function-based validation.** We implement a stratified P-fold version of our validation procedure as follows. First, we randomly split the training data into P mutually exclusive subsets, with $\mathcal{Z}_p$ being the set of indexes of data points in the p^{th} subset, and $\mathcal{Z}_{-p}$ its complement. In the p^{th} fold, the model being evaluated is trained on the data in $\mathcal{Z}_{-p}$, and issues a CATE estimate $\widehat{T}_{-p}$. For validation, we execute our two-step procedure as follows:

Step 1: Plug-in estimation (Section 3.1)

Using the dataset indexed by $\mathcal{Z}_{-p}$, we fit the plug-in model $\tilde{\theta}_{-p} = \{\tilde{\mu}_{-p,0}, \tilde{\mu}_{-p,1}, \tilde{\pi}_{-p}\}$ as explained in Section 3.1. We use two XGBoost regression models for $\tilde{\mu}_{-p,0}$ and $\tilde{\mu}_{-p,1}$, and then calculate $\widetilde{T}_{-p} = \tilde{\mu}_{-p,1} - \tilde{\mu}_{-p,0}$. For $\tilde{\pi}_{-p}$, we use an XGBoost classifier. Our choice of XGBoost is motivated by its minimax optimality (Linero & Yang, 2018), which is required by Theorem 1.

Step 2: Unplugged validation (Section 3.2)

Given $\tilde{\theta}_{-p}$, we estimate the model's PEHE on the held-out sample $\mathcal{Z}_p$ using the estimator in (9) with $m = 1$, i.e.,

$$\hat{\boldsymbol{\ell}}_p^{(1)} = \textstyle\sum_{i \in \mathcal{Z}_p} \Big[(\widehat{T}_{-p}(X_i) - \widetilde{T}_{-p}(X_i))^2 + \dot{\boldsymbol{\ell}}^{(1)}_{\tilde{\theta}_{-p}}(Z_i; \widehat{T}_{-p}) \Big],$$

where $\dot{\boldsymbol{\ell}}^{(1)}_{\theta}(.)$ is given by Theorem 2. (Here, the first order U-statistic $\mathbb{U}_1$ in (9) reduces to a sample average.)

The final PEHE estimate is given by the average PEHE estimates over the P validation folds, i.e., $\hat{\boldsymbol{\ell}}_n^{(1)} = n^{-1} \sum_p \hat{\boldsymbol{\ell}}_p^{(1)}$. In all experiments, we set $m = 1$ since higher order influence terms did not improve the results. This may be either because the condition $m \geq d/2(\alpha_0 \wedge \alpha_1)$ (Theorem 1) is satisfied with $m = 1$, or because we use approximate higher order influence (Appendix G). We defer investigations into the utility of higher order influence terms to future work.

Method name	Reference	% Winner
BNN★	Johansson et al. (2016)	3 %
CMGP‡	Alaa et al. (2017)	12 %
TARNet★	Shalit et al. (2017)	8 %
CFR Wass.★	Shalit et al. (2017)	12 %
CFR MMD★	Shalit et al. (2017)	9 %
NSGP★	Alaa et al. (2018)	17 %
GAN-ITE◇	Yoon et al. (2018)	7 %
SITE‡	Yao et al. (2018)	7 %
BART	Hill (2011)	15 %
Causal Forest	Wager et al. (2017)	10 %
Factual	—	53 %
IPTW	—	54 %
Plug-in	—	65 %
IF-based	—	72 %
Random	—	10 %
Supervised	—	84 %

Table 1. **Comparison of baselines over all datasets.** Refer to Appendix H for description of each method. (★ICML, ‡NeurIPS, ◇ICLR.)

■ **Automated causal inference.** Using our validation procedure, we validate a set of candidate models for a given dataset, and then pick the model with smallest $\hat{\boldsymbol{\ell}}_n^{(1)}$. Our candidate models include all methods published in ICML, NeurIPS and ICLR conferences from 2016 to 2018. This comprises a pool of 8 models, with modeling approaches ranging from Gaussian processes to generative adversarial networks. In addition, we included two other key models developed in the statistics community (BART and causal forests). All candidate models are presented in Table 1.

■ **Data description.** We conducted extensive experiments on benchmark datasets released by the "Atlantic Causal Inference Competition" (Hill, 2016), a data analysis competition that compared models of treatment effects. The competition involved 77 semi-synthetic datasets: all datasets shared the same data on features X, but each dataset had its own simulated outcomes and assignments (W, Y). Features were extracted from a real-world observational study, whereas outcomes and assignments were simulated via data generating processes that were carefully designed to mimic real data. Each of the 77 datasets had a unique data generating process encoding varying properties (e.g., levels of treatment effect heterogeneity, dimensionality of the relevant feature space, etc.) Detailed explanation of the data generating processes was published by the organizers of the competition in (Dorie et al., 2017).

The feature data shared by all datasets was extracted from the Collaborative Perinatal Project (Niswander, 1972), a study conducted on a cohort of pregnant women to identify causes of infants' developmental disorders. The treatment was a child's birth weight ($W = 1$ if weight $< 2.5\ kg$), and outcome was the child's IQ after a given follow-up period. The study contained 4,802 data points with 55 features (5 are binary, 27 are count data, and 23 are continuous).

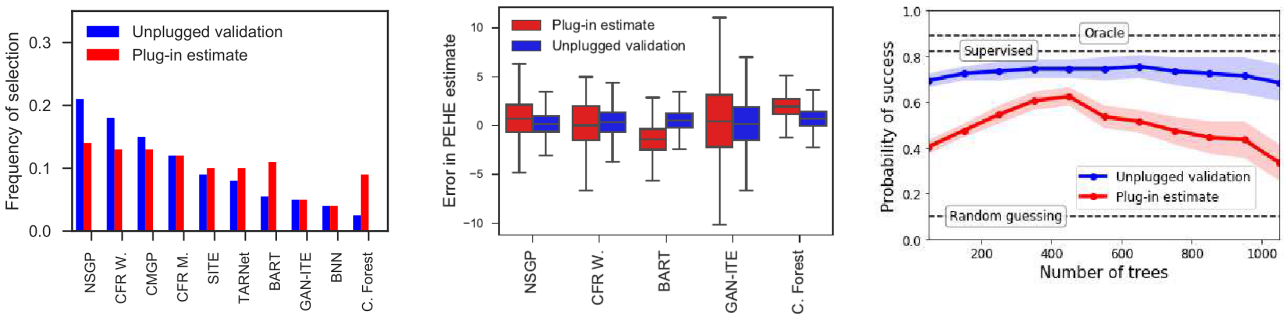

Figure 4. **Demonstration of the inner workings of influence function-based validation.** (Left) Frequency of selection of each model. (Middle) Box-plots for the errors in PEHE estimates for each model. (Right) Sensitivity to changes in the plug-in model.

■ **Performance evaluation.** We applied automated causal inference on 10 realizations of the simulated outcomes for each of the 77 datasets, i.e., a total of 770 replications. (Those realizations were generated by the competition organizers and are publicly accessible (Hill, 2016).) For each realization, we divide the data into 80/20 train/test splits, and use training data to predict the PEHE of the 10 candidate models via 5-fold influence function-based validation. Then, we select the model with smallest estimated PEHE, and evaluate its PEHE on the out-of-sample testing data.

■ **Baselines.** We compare influence function-based validation with 3 heuristics commonly used in the epidemiological and statistical literature (Schuler et al., 2018):

Baseline	PEHE estimator
Factual validation	$\hat{\ell}_n(\widehat{T}) = \frac{1}{n}\sum_i(\hat{\mu}_{W_i}(X_i) - Y_i^{(W_i)})^2$
IPTW validation	$\hat{\ell}_n(\widehat{T}) = \frac{1}{n}\sum_i \frac{(\hat{\mu}_{W_i}(X_i) - Y_i^{(W_i)})^2}{(1-2W_i)(1-W_i-\widetilde{\pi}(X_i))}$
Plug-in validation	$\hat{\ell}_n(\widehat{T}) = \frac{1}{n}\sum_i(\widehat{T}(X_i) - \widetilde{T}(X_i))^2$

Factual validation evaluates the error in the potential outcomes (μ_0, μ_1) using factual samples only. Inverse propensity weighted (IPTW) validation is similar, but weights each sample with its (estimated) "propensity score" $\widetilde{\pi}(x)$ to obtain unbiased estimates (Van der Laan et al., 2003). Plug-in validation is identical to Step 1 of our procedure: it obtains a plug-in PEHE estimate (Rolling & Yang, 2014; Schuler et al., 2017). To ensure fair comparisons, we model $\widetilde{T}$ and $\widetilde{\pi}$ in the heuristics above using XGBoost models similar to the ones used in Step 1 of our procedure.

5.2. Results and Discussion

Table 1 summarizes the fraction of datasets for which each baseline comes out as winner across all datasets[8]. As we can see, our influence function-based (IF-based) approach that automatically picks the best model for every dataset outperforms any single model applied repeatedly to all datasets. This is because the 77 datasets encode different data generating processes, and hence no single model is expected to be a good fit for all datasets. The gains achieved by automation are substantial — the PEHE of the automated approach was (on average) 47% smaller than that of the best performing single model.

It comes as no surprise that our procedure outperforms the factual, IPWT and plug-in validation heuristics. This is because, as we have shown in Theorem 1, the IF-based approach is the most efficient estimator of PEHE. We also compare our validation procedure with the "supervised" cross-validation procedure that is allowd to observe the counterfactual data in the training set. As we can see, despite having access to less information, our IF-based approach comes closer to the supervised approach (as compared to the competing validation methods).

In Figure 4, we trace the inner workings of our procedure by comparing the plug-in PEHE estimate $\boldsymbol{\ell}_{\tilde{\theta}}$ obtained in Step 1 with the corrected estimate $\hat{\boldsymbol{\ell}}_n^{(1)}$ obtained in Step 2, in terms of the frequency of selection of each model, the error in PEHE estimates per model, and the probability of successfully selecting the best model across the 77 datasets.

As we can see in Figure 4 (left), validation with the plug-in estimate $\boldsymbol{\ell}_{\tilde{\theta}}$ selects models almost arbitrarily (with nearly equal probabilities), but the corrected estimate $\hat{\boldsymbol{\ell}}_n^{(1)}$ is able to select well-performing ones more frequently. This is because, as shown in Figure 4 (middle), $\hat{\boldsymbol{\ell}}_n^{(1)}$ reduces the bias and variance incurred by the initial plug-in estimate, hence increasing the chance of distinguishing good models from bad ones. Figure 4 (right) shows that our procedure is robust to changes in the plug-in model — as we span the number of trees of the XGBoost-based plug-in model, we see little effect on the chance of selecting the best model. These results suggest that influence function-based validation can help practitioners leverage machine learning to enhance current practices in observational research.

[8] The magnitudes of causal effects were not consistent across datasets, hence PEHE values were in different numerical ranges.

Acknowledgements

The authors would like to thank the reviewers for their helpful comments. The research presented in this paper was supported by the Office of Naval Research (ONR) and the NSF (Grant number: ECCS1462245, ECCS1533983, and ECCS1407712).

References

Alaa, Ahmed and van der Schaar, Mihaela. Limits of estimating heterogeneous treatment effects: Guidelines for practical algorithm design. In *International Conference on Machine Learning*, pp. 129–138, 2018.

Alaa, Ahmed M and van der Schaar, Mihaela. Bayesian inference of individualized treatment effects using multi-task gaussian processes. *Advances in Neural Information Processing Systems (NIPS)*, 2017.

Atan, O, Jordon, J, and van der Schaar, M. Deep-treat: Learning optimal personalized treatments from observational data using neural networks. *AAAI*, 2018.

Basu, Ayanendranath, Harris, Ian R, Hjort, Nils L, and Jones, MC. Robust and efficient estimation by minimising a density power divergence. *Biometrika*, 85(3):549–559, 1998.

Bottou, Léon, Peters, Jonas, Quiñonero-Candela, Joaquin, Charles, Denis X, Chickering, D Max, Portugaly, Elon, Ray, Dipankar, Simard, Patrice, and Snelson, Ed. Counterfactual reasoning and learning systems: The example of computational advertising. *The Journal of Machine Learning Research*, 14(1):3207–3260, 2013.

Chernozhukov, Victor, Chetverikov, Denis, Demirer, Mert, Duflo, Esther, Hansen, Christian, Newey, Whitney, and Robins, James. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

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. *arXiv preprint arXiv:1707.02641*, 2017.

Dudoit, Sandrine and van der Laan, Mark J. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2 (2):131–154, 2005.

Fernholz, Luisa Turrin. *Von Mises calculus for statistical functionals*, volume 19. Springer Science & Business Media, 2012.

Foster, Jared C, Taylor, Jeremy MG, and Ruberg, Stephen J. Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880, 2011.

Franklin, Jessica M, Schneeweiss, Sebastian, Polinski, Jennifer M, and Rassen, Jeremy A. Plasmode simulation for the evaluation of pharmacoepidemiologic methods in complex healthcare databases. *Computational statistics & data analysis*, 72:219–226, 2014.

Goldstein, Larry and Messer, Karen. Optimal plug-in estimators for nonparametric functional estimation. *The annals of statistics*, pp. 1306–1328, 1992.

Hahn, P Richard, Murray, Jared S, and Carvalho, Carlos M. Bayesian regression tree models for causal inference: regularization, confounding, and heterogeneous effects. 2017.

Hampel, Frank R, Ronchetti, Elvezio M, Rousseeuw, Peter J, and Stahel, Werner A. *Robust statistics: the approach based on influence functions*, volume 196. John Wiley & Sons, 2011.

Hill, Jennifer L. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011.

Hill, Jennifer L. 2016 atlantic causal inference conference competition: Is your satt where it's at?, 2016. URL http://jenniferhill7.wixsite.com/acic-2016/competition.

Johansson, Fredrik, Shalit, Uri, and Sontag, David. Learning representations for counterfactual inference. In *International Conference on Machine Learning*, pp. 3020–3029, 2016.

Johansson, Fredrik D, Kallus, Nathan, Shalit, Uri, and Sontag, David. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*, 2018.

Kallus, Nathan. Recursive partitioning for personalization using observational data. In *International Conference on Machine Learning*, pp. 1789–1798, 2017.

Kennedy, Edward H. Nonparametric causal effects based on incremental propensity score interventions. *Journal of the American Statistical Association*, 2018.

Koh, Pang Wei and Liang, Percy. Understanding black-box predictions via influence functions. In *International Conference on Machine Learning*, pp. 1885–1894, 2017.

Künzel, Sören, Sekhon, Jasjeet, Bickel, Peter, and Yu, Bin. Meta-learners for estimating heterogeneous treatment effects using machine learning. *arXiv preprint arXiv:1706.03461*, 2017.

Li, Sheng and Fu, Yun. Matching on balanced nonlinear representations for treatment effects estimation. In *Advances in Neural Information Processing Systems*, pp. 930–940, 2017.

Linero, Antonio R and Yang, Yun. Bayesian regression tree ensembles that adapt to smoothness and sparsity. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(5):1087–1110, 2018.

Longford, Nicholas T. A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74(4):817–827, 1987.

Niswander, Kenneth R. The collaborative perinatal study of the national institute of neurological diseases and stroke. *The Woman and Their Pregnancies*, 1972.

Oliver, Avital, Odena, Augustus, Raffel, Colin, Cubuk, Ekin D, and Goodfellow, Ian J. Realistic evaluation of semi-supervised learning algorithms. *International Conference on Learning Representations (ICLR)*, 2018.

Pearl, Judea. *Causality*. Cambridge university press, 2009.

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. *Statistics in medicine*, 37(11):1767–1787, 2018.

Ray, Kolyan and van der Vaart, Aad. Semiparametric bayesian causal inference using gaussian process priors. *arXiv preprint arXiv:1808.04246*, 2018.

Robins, James, Li, Lingling, Tchetgen, Eric, van der Vaart, Aad, et al. Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, pp. 335–421. Institute of Mathematical Statistics, 2008.

Robins, James M, Li, Lingling, Mukherjee, Rajarshi, Tchetgen, Eric Tchetgen, van der Vaart, Aad, et al. Minimax estimation of a functional on a structured high-dimensional model. *The Annals of Statistics*, 45(5): 1951–1987, 2017.

Rolling, Craig A and Yang, Yuhong. Model selection for estimating treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4): 749–769, 2014.

Rubin, Donald B. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

Rubin, Donald B. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

Rubin, Donald B. On the limitations of comparative effectiveness research. *Statistics in medicine*, 29(19):1991–1995, 2010.

Schuler, Alejandro, Jung, Ken, Tibshirani, Robert, Hastie, Trevor, and Shah, Nigam. Synth-validation: Selecting the best causal inference method for a given dataset. *arXiv preprint arXiv:1711.00083*, 2017.

Schuler, Alejandro, Baiocchi, Michael, Tibshirani, Robert, and Shah, Nigam. A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146*, 2018.

Setoguchi, Soko, Schneeweiss, Sebastian, Brookhart, M Alan, Glynn, Robert J, and Cook, E Francis. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiology and drug safety*, 17(6):546–555, 2008.

Shalit, Uri, Johansson, Fredrik, and Sontag, David. Estimating individual treatment effect: generalization bounds and algorithms. pp. 3076–3085, 2017.

Shimodaira, Hidetoshi. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90 (2):227–244, 2000.

Shmueli, Galit et al. To explain or to predict? *Statistical science*, 25(3):289–310, 2010.

Stuart, Elizabeth A, DuGoff, Eva, Abrams, Michael, Salkever, David, and Steinwachs, Donald. Estimating causal effects in observational studies using electronic health data: challenges and (some) solutions. *Egems*, 1 (3), 2013.

Subbaswamy, Adarsh and Saria, Suchi. Counterfactual normalization: Proactively addressing dataset shift using causal mechanisms. In *Uncertainty in Artificial Intelligence*, pp. 947–957, 2018.

Van der Laan, Mark J, Laan, MJ, and Robins, James M. *Unified methods for censored longitudinal data and causality*. Springer Science & Business Media, 2003.

van der Vaart, Aad. Higher order tangent spaces and influence functions. *Statistical Science*, pp. 679–686, 2014.

Wager, Stefan and Athey, Susan. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

Wright, Daniel B, London, Kamala, and Field, Andy P. Using bootstrap estimation and the plug-in principle for clinical psychology data. *Journal of Experimental Psychopathology*, 2(2):jep–013611, 2011.

Xie, Yu, Brand, Jennie E, and Jann, Ben. Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1):314–347, 2012.

Yao, Liuyi, Li, Sheng, Li, Yaliang, Huai, Mengdi, Gao, Jing, and Zhang, Aidong. Representation learning for treatment effect estimation from observational data. In *Advances in Neural Information Processing Systems*, pp. 2634–2644, 2018.

Yoon, Jinsung, Jordon, James, and van der Schaar, Mihaela. Ganite: Estimation of individualized treatment effects using generative adversarial nets. *International Conference on Learning Representations (ICLR)*, 2018.

Zhao, Qingyuan et al. Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics*, 47 (2):965–993, 2019.