In Search of Insights, Not Magic Bullets: Towards Demystification of the Model Selection Dilemma in Heterogeneous Treatment Effect Estimation

Alicia Curth [1] **Mihaela van der Schaar** [1 2]

Abstract

Personalized treatment effect estimates are often of interest in high-stakes applications – thus, before deploying a model estimating such effects in practice, one needs to be sure that the best candidate from the ever-growing machine learning toolbox for this task was chosen. Unfortunately, due to the absence of counterfactual information in practice, it is usually not possible to rely on standard validation metrics for doing so, leading to a well-known model selection dilemma in the treatment effect estimation literature. While some solutions have recently been investigated, systematic understanding of the strengths and weaknesses of different model selection criteria is still lacking. In this paper, instead of attempting to declare a global 'winner', we therefore empirically investigate success- *and* failure modes of different selection criteria. We highlight that there is a complex interplay between selection strategies, candidate estimators and the data used for comparing them, and provide interesting insights into the relative (dis)advantages of different criteria alongside desiderata for the design of further illuminating empirical studies in this context.

1. Introduction

Applications in which the causal effects of treatments (or actions, interventions & policies) are of interest are ubiquitous in empirical science, and personalized effect estimates could ultimately be used to improve decision making in many domains, including healthcare, economics and marketing. Machine learning (ML) has shown great promise in providing such personalized effect estimates (Bica et al., 2021), and the ML literature on the topic has matured over the last five years: a plethora of new methods for estimating conditional average treatment effects (CATE) have been proposed recently, including *method-agnostic* so-called meta-learner strategies that could be implemented using any ML prediction method (Künzel et al., 2019; Nie & Wager, 2021; Kennedy, 2020; Curth & van der Schaar, 2021b) as well as adaptations of specific ML methods (Shalit et al., 2017; Alaa & van der Schaar, 2018; Wager & Athey, 2018) to the treatment effect estimation context.

Personalized treatment effect estimates, crucially, are often of interest in safety-critical applications, particularly in medicine and policy making. Thus, prior to use in practice we would like to ensure that we have selected a ML estimator from this vast toolbox that is *trustworthy* and outputs the best possible estimates. While this sounds like a straightforward requirement, it remains a big hurdle in practice: because of the *fundamental problem of causal inference* (Holland, 1986), ground truth treatment effects are usually not available to perform standard model validation, and alternative solutions need to be considered to overcome this *model selection dilemma*. As we discuss below, despite recent proposals of new model selection strategies and recent empirical studies comparing different strategies, we believe that there is still a general lack of understanding of the (dis)advantages of different strategies and how they are entangled with underlying data-generating processes (DGPs) – which we aim to provide in this work.

Related work. Despite its practical relevance, the problem of model selection for heterogeneous treatment effect estimation has received only very limited attention so far (this stands also in stark contrast to the plethora of new estimators proposed in recent years, see Appendix B for an overview). An intuitive and often applied solution is to rely on a simple prediction-type validation and evaluate a model's performance in *predicting the observed (factual) outcome* associated with the factual treatment *instead* of evaluating the quality of the *effect estimate*. More targeted alternatives have recently been developed: Rolling & Yang (2014) propose to construct approximate effect validation targets by matching the nearest treated and control units and comparing their outcomes, Nie & Wager (2021) highlight that their R-learner objective could also be used for

[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, UK [2]The Alan Turing Institute. Correspondence to: Alicia Curth <amc253@cam.ac.uk>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

model selection, Saito & Yasui (2020) similarly propose the use of a criterion that corresponds to Kennedy (2020)'s meta-learner objective and Alaa & Van Der Schaar (2019) propose to rely on influence functions to de-biase plugin estimates. We are aware of two independent benchmarking studies that compare (subsets of) such criteria: Schuler et al. (2018) find that the R-learner objective performs best overall, while Mahajan et al. (2022) find that no criterion dominates all others over all datasets considered (and in particular do not find the R-learner objective to perform remarkably), which highlights to us that there is much room for understanding of the relative strengths but also relative weaknesses of different selection criteria.

Contributions. In this paper, we focus on building *systematic understanding* of the (dis)advantages of different model selection criteria. Note that therefore our aim is not to propose new methodology, but rather to establish understanding and insight into the tools already available in the literature. We believe that this is one of the most crucial and necessary next steps for this community in order to enable actual adoption of personalized treatment effect estimators in practice, and may inspire further methodological research to fill gaps highlighted by this understanding. In doing so, we make three contributions:

1. We develop intuition for a highly complex selection problem: we shine light on its inherent challenges, provide structure to existing work by presenting a classification of existing criteria, and use these insights to derive hypotheses for their relative performance.
2. We present desiderata for experimental design that enable us to disentangle the complex forces at play in this problem: we advocate for better experiments that allow to systematically investigate the interplay between DGP, candidate estimators and selection criteria through reliance on data-generating processes with interesting axes of variation and more transparent reporting practices.
3. We provide new insights into the CATE model selection problem through an empirical investigation of the success and failure modes of existing criteria, and conclude that no existing selection criterion is globally best across all experimental conditions we consider. Next to highlighting some performance trends across the different types of selection criteria, we mainly focus on investigating i) congeniality biases between candidate estimators and selection criteria imbued with similar inductive biases in their construction and ii) what factual selection criteria can(not) achieve. We find that i) selection criteria relying on plug-in estimates of treatment effects are likely to favor estimators that resemble their plug-ins, while in selection criteria relying on pseudo-outcomes such congeniality biases are less pronounced, and that ii) factual selection sometimes underperforms not only because it cannot evaluate all types of CATE estimators, but also because it is not well-targeted at effect estimates.

2. Problem Setting

We consider the by now standard CATE estimation setup within the potential outcomes framework (Rubin, 2005). That is, we assume access to a dataset consisting of n i.i.d. tuples (X, A, Y), where Y is an outcome of interest, X consists of *pre-treatment* covariates and A 2 f0; 1g is a binary treatment (action, intervention or policy), which is assigned according to some (often unknown) propensity $\pi(x) = \mathsf{P}(A = 1 | X = x)$. We assume that conceptually each individual is a priori associated with two *potential outcomes (POs)* $Y(a), a$ 2 f0; 1g, capturing outcome under either treatment a, however, we observe only the outcome associated to the treatment A actually received, i.e. $Y = Y(A)$, We can thus naturally define an individualized treatment effect through the (unobserved) PO contrast $Y(1)$ $Y(0)$. We focus on estimating the conditional average treatment effect (CATE) $\tau(x)$, i.e. the expected PO difference for an individual with covariates $X = x$:

$$\tau(x) = \mathsf{E}[Y(1) \quad Y(0) | X = x] = \mu_1(x) \quad \mu_0(x) \quad (1)$$

where $\mu_a(x) = \mathsf{E}[Y(a) | X = x]$. To ensure that effects are identifiable and nonparametrically estimable from observational data, we rely on the standard ignorability assumptions (Rosenbaum & Rubin, 1983):

Assumption 2.1 (Ignorability.). (i) *Consistency.* For an individual with treatment assignment A, we observe the associated potential outcome, i.e. $Y = Y(A)$. (ii) *Unconfoundedness.* There are no unobserved confounders, so that $Y(0), Y(1) \perp\!\!\!\perp A | X$. (iii) *Overlap.* Treatment assignment is non-deterministic, i.e. $\pi(x)$ 2 $(0, 1)$.

Then, $\mathsf{E}[Y(a) | X = x] = \mathsf{E}[Y | X = x, A = a]$ so that observed statistical associations have a causal interpretation.

2.1. CATE Estimation Strategies

A plethora of strategies for estimating CATE have been proposed in the recent literature. One strand of this work has recently relied on the *meta-learner* framework of Künzel et al. (2019), where a meta-learner provides a 'recipe' for estimating CATE using *any arbitrary*[1] ML method M. Due to their ease of implementation with different underlying ML methods, existing theoretical understanding and correspondence to the model selection strategies discussed in the following section, we focus on the problem of choosing between such meta-learners in this paper.

Following Curth & van der Schaar (2021b) we distinguish between (i) *indirect estimation strategies*, which estimate

[1] as opposed to *adaptation of specific* ML methods proposed in another stream of work, as discussed further in Appendix B.

CATE *indirectly* by outputting estimates $\hat{\mu}_a(x)$ of the PO regressions and then setting $\hat{\tau} = \hat{\mu}_1(x) \quad \hat{\mu}_0(x)$, and (ii) *direct estimation strategies* which output an estimate $\hat{\tau}(x)$ *directly* without outputting PO estimates as byproducts. Künzel et al. (2019) discuss two *indirect learners*: a T-learner strategy, where the training data is split by treatment group and M is trained independently (twice) on each sample to output two regressors $f\hat{\mu}_0(x), \hat{\mu}_1(x)g$, and a S-learner strategy, where the treatment indicator A is simply appended to X so that M can be trained a single time using covariates (X, A) and outputting a single estimated function $\hat{\mu}(x, a)$ that can be used to impute both POs. Because the latter formulation can lead to implicit regularization of CATE (Schuler et al., 2018) (as any heterogeneous effect has to be represented by *learned interaction terms* of X and A), we also include an extended version (ES-learner) which is trained on the covariates $(X, X \quad A, A)$ explicitly.

Alternatively, there exist *direct estimators* that output an estimate of the CATE *only* by relying on estimates of (some of) the nuisance parameters $\eta = (\mu_0(x), \mu_1(x), \pi(x))$ obtained in a pre-processing step using ML method M. Most such strategies, in particular, X-learner (Künzel et al., 2019), the DR-learner (Kennedy, 2020) and RA- and PW-learner (Curth & van der Schaar, 2021b), rely on a pseudo-outcome approach where, using plug-in nuisance estimates $\hat{\eta}$, one constructs a pseudo outcome for which it holds that $\mathsf{E}[Y_\eta | X = x] = \tau(x)$ for ground truth nuisance parameters η, and then regresses $Y_{\hat{\eta}}$ on X using M to give an estimate $\hat{\tau}(x)$. Nie & Wager (2021)'s R-learner is similar in spirit but relies on a modified loss function instead. These multi-stage estimation procedures have recently gained popularity in the literature because they have good theoretical properties (Curth & van der Schaar, 2021b), are more robust (Kennedy, 2020) and have been observed to perform better across a variety of scenarios than vanilla S- and T-learner individually (Nie & Wager, 2021).

3. CATE Model Selection: Understanding Challenges and Existing Strategies

We study the problem of selecting an estimator from a set of CATE estimators $\mathcal{T} = f\hat{\tau}_1(\ ), \ldots, \hat{\tau}_K(\ )g$, containing different meta-learner+ML method combinations, that minimizes the precision of estimating heterogeneous effects (PEHE) (Hill, 2011), the root-mean-squared error of estimating the underlying effect function over a test-set of size n:

$$\arg\min_{\hat{\tau}_k \in \mathcal{T}} \mathsf{E}_\tau^{oracle}(\hat{\tau}_k) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tau(X_i) \quad \hat{\tau}_k(X_i))^2}$$

This metric is an *oracle* metric because it cannot be evaluated in practice – making the CATE model selection problem highly nontrivial. Below, we therefore provide an in-depth discussion of the inherent challenges of this problem, and then establish a classification of model selection criteria that have been proposed to overcome these challenges.

3.1. What makes CATE model selection challenging?

Challenge 1: Lack of supervised signal for the individual treatment effect. Due to the *fundamental problem of causal inference* (Holland, 1986), i.e. the fact that we can only *either* observe $Y(0)$ or $Y(1)$ for any one individual, the true supervised target label $Y(1) \quad Y(0)$ for estimation of $\mathsf{E}[Y(1) \quad Y(0) | X = x] = \tau(x)$ is not available for model selection through a standard held-out validation approach. This lack of supervised label is also the issue that motivated the construction of the direct meta-learners for *estimation* of effects using pseudo-outcomes (Kennedy, 2020; Curth & van der Schaar, 2021b), but it results in another challenge for model selection: even though direct estimation of effects is possible, it is not possible to compare multiple direct estimators on basis of their output $\hat{\tau}_k(x)$ using *observed data only* because there is no natural outcome to validate $\hat{\tau}_k(x)$ against directly.

Challenge 2: Confounding leads to covariate shift between treatment groups. One straightforward option to validate estimators using factual (observed) outcomes only is to simply evaluate them based on their outcome prediction ability; i.e. to use $Y_i \quad \hat{\mu}_{A_i}(X_i)$ for validation. This, however, is only an option for evaluating indirect learners because it requires an output $\hat{\mu}_a(x)$. Even when one is willing to restrict attention only to indirect learners to make use of a factual evaluation strategy, a remaining inherent challenge is that evaluating $\hat{\mu}_a(x)$ only on individuals with $A = a$ observed inherently suffers from *covariate shift* whenever $\pi(x)$ is not constant because treatment is not assigned completely randomly. This too is a challenge also when estimating effects (Shalit et al., 2017).

Challenge 3: Selection for good PO estimation and CATE estimation may not be the same. Finally, even when selecting only among indirect learners and in absence of covariate shift, in finite samples the estimator with the best performance on estimating (potential) outcomes might not do best at estimating CATE. It is clear that when models are correctly specified and unlimited data is available, perfectly estimating the POs will immediately lead to perfect CATE estimates. However, when data is limited and/or the model is misspecified, this might lead to a trade-off between estimating the POs well and estimating their difference[2].

[2]To see this, note that when comparing an indirect estimator $\hat{\mu}_a^1(x) = f\mu_1(x) + \epsilon_1(x), \mu_1(x) \quad \epsilon_1(x)g$ to an estimator $\hat{\mu}_a^2(x) = f\mu_1(x) + \epsilon_2(x), \mu_1(x) + \epsilon_2(x)g$ with estimation errors $\epsilon_1(x) < \epsilon_2(x)$ for all x, the MSE of estimating the POs will be lower for estimator $\hat{\mu}_a^1(x)$ – yet its estimation error on CATE will be $2\epsilon_1(x)$ for every x while estimator $\hat{\mu}_a^2(x)$ with worse estimation on the POs will have CATE estimation error 0.

3.2. Categorizing model selection criteria

To overcome the challenges discussed above, numerous alternative model selection criteria have been used or proposed in related work. Below, we establish a classification of existing model selection criteria into three categories based on their most salient characteristics: we consider factual (prediction) criteria, plug-in surrogate criteria and pseudo-outcome surrogate criteria. We provide a conceptual overview of the three types of strategies in Fig. 1. Overall, we consider a similar set of model selection criteria as the (union of) the benchmarking studies presented in Schuler et al. (2018); Mahajan et al. (2022)[3].

1. Factual (prediction) criteria. First, as discussed in the previous section, a possible way of evaluating models $\hat{\tau}_k(x)$ that also output a pair of regressors $f\hat{\mu}_0^k(x), \hat{\mu}_1^k(x)g$ is to rely on a simple prediction loss considering only the observed potential outcome

$$E_Y^{fact}(\hat{\tau}_k) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i \quad \hat{\mu}_{A_i}^k(X_i))^2}$$

In order to correct for possible covariate shift, this can also be transformed into an importance weighted criterion $E_Y^{fact,w}$ using a propensity score estimate in $w(X_i, A_i) = A_i(\hat{\pi}(X_i)) \ ^1 + (1 \quad A_i)(1 \quad \hat{\pi}(X_i)) \ ^1$.

2. Plug-in surrogate criteria. To actually evaluate estimates $\hat{\tau}_k(x)$ *directly* one is thus forced to construct surrogates for CATE. One way of doing so is by fitting a new CATE estimator $\tilde{\tau}(x)$ on held-out data and using this to compare against the estimates:

$$E_\tau^{plug}(\hat{\tau}_k) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\tilde{\tau}(X_i) \quad \hat{\tau}_k(X_i))^2}$$

For implementation of this criterion, any CATE estimator could be used – thus trying to select the best plug-in estimator may potentially lead us in a circle and back to the problem we are originally trying to solve. Related work (Alaa & Van Der Schaar, 2019; Mahajan et al., 2022) considered only indirect estimators (S and T-learners) as a plug-in surrogate criterion – possibly because it is possible to choose between those factually – but we note that technically any estimator, including direct ones, could be used as $\tilde{\tau}(x)$.

[3]Schuler et al. (2018), predating publication of most model selection papers, miss some of the plug-in and pseudo-outcome criteria. Mahajan et al. (2022) only do not consider factual (prediction) criteria, which we consider of major interest as discussed in the following section. We only drop *policy value* criteria, i.e. those optimizing the derived treatment policy $\mathbf{1}f\hat{\tau}_k(x) > 0g$, from consideration, both because we focus on PEHE and because these criteria were shown to underperform *even when evaluated in terms of policy value* (Schuler et al., 2018).

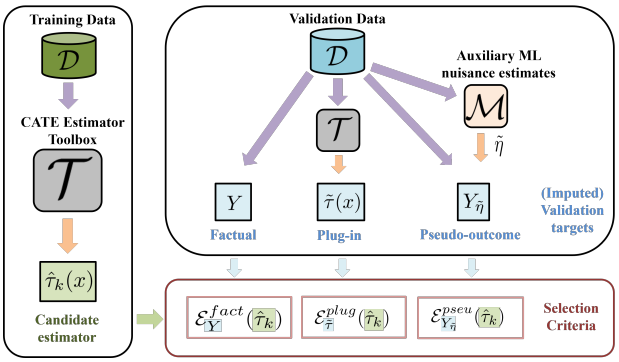


Figure 1. Conceptual overview of the considered selection criteria.

3. Pseudo-outcome surrogate criteria. Finally, one could make use of the same pseudo-outcome based strategy that underlies the direct learners: given auxiliary nuisance estimates $\tilde{\eta} = (\tilde{\mu}_0(x), \tilde{\mu}_1(x), \tilde{\pi}(x))$ obtained from the validation data using ML method M, one can construct pseudo-outcomes Y_η for which it holds that for ground truth nuisance parameter η, $E[Y_\eta jX = x] = \tau(x)$ and – instead of using them as regression outcomes as in the learners themselves – one can use them as validation targets in

$$E_{Y_\eta}^{pseu}(\hat{\tau}_k) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_\eta \quad \hat{\tau}_k(X_i))^2}$$

which is reasonable because the conditional mean $E[Y_\eta jX = x]$ minimizes the MSE $E_{Y_\eta}^{pseu}(\ )^2$ in expectation. Use of the doubly robust pseudo-outcome,

$$Y_{DR,\eta} = \left(\frac{A}{\hat{\pi}(X)} \quad \frac{(1 \quad A)}{1 \quad \hat{\pi}(X)}\right) Y + \left[\left(1 \quad \frac{A}{\hat{\pi}(X)}\right)\hat{\mu}_1(x) \quad \left(1 \quad \frac{1 \quad A}{1 \quad \hat{\pi}(X)}\right)\hat{\mu}_0(X)\right]$$

which is what the direct meta-learner known as the DR-learner (Kennedy, 2020) is based on, gives rise to the selection criterion proposed in Saito & Yasui (2020). We note here that it would also be possible to use the other meta-learner pseudo-outcomes discussed in Curth & van der Schaar (2021b) for this purpose, e.g. the singly-robust propensity-weighted $Y_{PW,\eta} = \left(\frac{A}{\hat{\pi}(X)} \quad \frac{(1 \quad A)}{1 \quad \hat{\pi}(X)}\right) Y$. We also consider Rolling & Yang (2014)'s matching based model selection strategy to fall into this category. In this case $Y_{\text{match},\eta} = (2A_i \quad 1)(Y_i \quad \tilde{NN}_{1 \quad A_i}(X_i))$ where $\tilde{NN}_a(X)$ is the nearest neighbor of X in treatment group a; this essentially corresponds to the pseudo-outcome associated with the RA-learner of Curth & van der Schaar (2021b), implemented using 1-NN regression to estimate the nuisance parameters $\hat{\mu}_a(x)$. We also put Alaa & Van Der Schaar (2019)'s influence function based criterion, which we discuss in Appendix C, into this category. Finally, the R-learner objective of Nie & Wager (2021), which requires an estimate

of the treatment-unconditional mean $\mu(x) = \mathsf{E}[YjX = x]$, relies on a similar idea[4] and can also been used for the selection task (Nie & Wager, 2021), resulting in the criterion

$$E_R^{pseu}(\hat{\tau}_k) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i \quad \tilde{\mu}(X_i) \quad \hat{\tau}_k(X_i)(A_i \quad \tilde{\pi}(X_i)))^2}$$

4. Demystifying the Model Selection Dilemma

Having established a high-level classification of model selection strategies, we can now build intuition and hypothesize about their expected (dis)advantages based on their inherent characteristics (Sec. 4.1). These expectations lead to numerous interesting research questions, which we believe can only be disentangled by designing carefully controlled experiments – a feature that we would argue related work has neglected so far. Therefore, we then move to discuss design principles for construction of empirical studies of the CATE model selection problem (Sec. 4.2), which we will then apply in our experiment section.

4.1. Comparing model selection criteria: expectations on advantages and disadvantages

Combining on our high-level overview of different strategies for CATE model selection with the previously discussed challenges, we argue that every *class* of criteria comes with their own inherent (dis)advantages.

Factual (prediction) criteria. We believe that factual criteria could be very appealing for use in practice because – at least the unweighted E_Y^{fact} – does not require estimates of any nuisance parameters and *only relies on observed data*; this means that there is no additional overhead and its results could be considered trustworthy in the sense that there is no dependence on possibly misspecified or biased nuisance estimates. However, such criteria mainly evaluate the performance in terms of estimation of the POs, which, as discussed above, may wrongly prioritize good fit on the POs over good CATE fit – while the latter cannot be measured at all. This last point, crucially, also means that the factual criterion *cannot* evaluate all types of methods, excluding all direct estimators in particular, and therefore has to select among a smaller set $T_{indirect} \quad T$ of all possible estimators – potentially missing out on the estimators with the best performance by construction.

Plug-in surrogate criteria. Plug-in surrogate criteria have the clear advantage over factual criteria that they can evaluate *all* types of estimators and are targeted at the outcome of interest (i.e. the CATE). Yet, because a plug-in estimate $\tilde{\tau}(x)$ is needed, this introduces additional potential for estimation error or variance. Further, $\tilde{\tau}(x)$ could be any CATE estimator, thus choosing a good plugin $\tilde{\tau}(x)$ leads us back to the dilemma we were initially trying to overcome. Finally, we believe that such surrogate criteria may also suffer from a phenomenon we will refer to as ***congeniality bias***: they may advantage CATE estimators $\hat{\tau}_k(x)$ that are *structurally similar* to their plug-in estimator $\tilde{\tau}(x)$. Even though $\hat{\tau}_k(x)$ and $\tilde{\tau}(x)$ should be fit on different data folds, we expect that a plug-in criterion $\tilde{\tau}(x)$ may still prefer estimators imbued with similar inductive biases. That is, we expect that e.g. a criterion using plug-in surrogate $\tilde{\tau}(x)$ implemented using linear regression may favor CATE estimators $\hat{\tau}_k(x)$ similarly relying on linear regressions (over estimators implemented using other methods M), and one relying on an S-learner surrogate $\tilde{\tau}(x)$ may have a preference for selecting S-learners $\hat{\tau}_k(x)$. Here, we borrow the term 'congeniality bias' from the psychology literature, where it is used to indicate that individuals may have a systematic preference for information consistent with current beliefs (Hart et al., 2009).

Pseudo-outcome surrogate criteria. We expect these criteria to share the advantages, and some of the disadvantages of the plug-in criteria (namely the need to estimate additional parameters, and resulting possibility for increased error or variance). Because they do not use a final $\tau(x)$ estimate but only a pseudo-outcome Y_η, we expect that they might be *less* likely to suffer from congeniality bias, but could still favor estimators with similar inductive biases, e.g. direct estimators trained on the same pseudo-outcome.

4.1.1. Resulting research questions.

This paper was motivated by many of the interesting research questions outlined below that naturally follow from the discussion above, none of which we believe have been addressed in related work– we believe that this is a result of the fact that the focus so far has been on establishing *global best performance* of some criterion. Instead, we are interested in understanding scenarios in which there could be performance differences of competing criteria, in the hope that this will help practitioners in choosing the right criterion in their specific application. In particular, we are interested in exploring three questions in depth in this paper:

- **Q1.** When do which selection criteria work better or worse? If there are systematic patterns, do they parallel those observed for the different *estimation strategies* in prior work?
- **Q2.** Do surrogate selection criteria truly suffer from congeniality bias as expected? Are there differences between the different types?
- **Q3.** What and when do we lose out on by relying on factual validation $E_{\mu_a}^{fact}$? Does it matter that we restrict the estimator pool? Does it matter that we are optimizing for the wrong target?

[4] E_R^{pseu} can be rewritten in pseudo-outcome form as $Y_{R,\eta} = \frac{Y_i \quad \pi(X_i)}{A_i \quad \pi(X_i)}$ if combined with weights $(A_i \quad \pi(X_i))^2$ to be used in the sum inside the RMSE (Knaus et al., 2021).

4.2. Establishing desiderata for experimental design

As outlined above, we believe that there are many very interesting questions to explore in this CATE model selection dilemma. Existing empirical studies that we are aware of – both those proposing new model selection criteria (Alaa & Van Der Schaar, 2019; Saito & Yasui, 2020) and those benchmarking existing ones (Schuler et al., 2018; Mahajan et al., 2022) – have mainly focused on the question *'what strategy works best globally?'*, and have taken a black-box approach in doing so: by considering opaque datasets for benchmarking, by considering a large inextricable set of estimators to select from and by mainly reporting on averages across a number of different DGPs.

It has recently been highlighted that the benchmarking practices in the ML CATE estimation literature more generally have many shortcomings (Curth et al., 2021), especially the reliance on single semi-synthetic datasets[5] that encode very specific problem characteristics in their DGP, without discussing the effect of these choices. We therefore believe that it is crucial to carefully design controlled experimental environments that allow to disambiguate the effects that different components of a DGP may have on selection criterion performance. Below we discuss desiderata for designing an empirical study that allows for the insight into model selection performance we seek, which we use to design an empirical study in the following section.

1. Use DGPs with interesting 'experimental knobs'. To gain systematic understanding of *when* different selection criteria work well, we need to be able to *systematically* vary the underlying experimental characteristics – as suggested by Dorie et al. (2019) we therefore design simulations that enable turning of important 'experimental knobs'. In pursuit of interesting insights into relative performance of selection criteria, we will therefore choose to investigate axes that have been shown to matter for estimator performance itself as well as others that we would expect to matter.

2. Examine the performance of candidate estimators in $\mathcal{T}$. To understand when different selectors perform well, we believe that it is important to first establish how the underlying candidate estimators in $\mathcal{T}$ perform. All related work that we are aware of skip this step, yet we consider it crucial because the performance of different selection strategies may be deeply entangled with the performance of the underlying estimators: if, for example, there is congeniality between selectors and estimators, and/or a specific type of estimator is advantaged on a specific dataset, then the corresponding selector may perform well by construction.

[5]Note that, due to the lack of $Y(1)$ $Y(0)$ in real data it is generally *necessary to simulate* outcomes in experiments to allow for known ground truth; most existing benchmark datasets are *semi-synthetic* in that they use real covariates but simulate outcomes.

3. Analyze how and *when* performance of selection criteria varies. Finally, once DGPs are constructed and estimators examined, we aim to analyze the performance of the selection criteria in detail. That is, we wish to explicitly understand how the relative performance of selectors varies as an experimental knob is turned. In many related works this is not possible as results are reported as *averages* across many different settings (Alaa & Van Der Schaar, 2019; Mahajan et al., 2022), obfuscating possible interesting insights into systematic performance differences.

5. Empirical Study

Setup. In this section, we conduct an empirical study comparing CATE selection criteria following the three steps outlined above. Throughout, we rely on two ML-methods M to instantiate all meta-learners and selection criteria: Extreme Gradient Boosted Trees (Chen & Guestrin, 2016) (GB) and linear regressions with ridge penalty (LR). We chose these two because they encode very different inductive biases, and allow us to give insights into performance differences between very flexible versus rigid models. Whenever propensity score estimates are needed, we estimate these using logistic regressions. As meta-learners to choose between, we consider (indirect) S-, T- and ES-learners and the (direct) DR- and R-learner. As selection criteria we consider E_Y^{fact} (factual), $E_\tau^{plug,ES}$, $E_\tau^{plug,T}$, $E_\tau^{plug,DR}$ & $E_\tau^{plug,R}$ (surrogate plug-in) and $E_{Y_\eta}^{pseudo,DR}$ & $E_{Y_\eta}^{pseudo,R}$ (surrogate pseudo-outcome) in the main text, further results using criteria that generally performed worse can be found in Appendix D.1. Implementation details[6] can be found in Appendix C.

5.1. Step 1: Designing an insightful DGP.

We build a DGP loosely inspired by the setup used in Curth & van der Schaar (2021a): we similarly make use of the covariates X of the ACIC2016 dataset (Dorie et al., 2019), and simulate our own outcomes and treatment assignments for greater transparency and control. We begin by binarizing all continuous covariates at randomly sampled cutoff points, obtaining processed covariates X , and then use them in a linear model for $\mu_0(\,)$, including up to third-order interaction terms of X , while $\tau(\,)$ is simply linear in X . This setup includes 3 main experimental knobs:

1. **CATE complexity.** Our first experimental knob is $\rho \in \{0, .1, .3\}$, the proportion of non-zero coefficients of inputs in $\tau(\,)$, controlling the complexity (sparsity) of $\tau(\,)$ – this was used in Curth & van der Schaar (2021a) and shown to matter for relative performance of *estimators*.
2. **Misspecification.** Second, we introduced a transformation of covariates deliberately because it allows us to

[6]Code to replicate all experiments is available at `https://github.com/AliciaCurth/CATESelection`

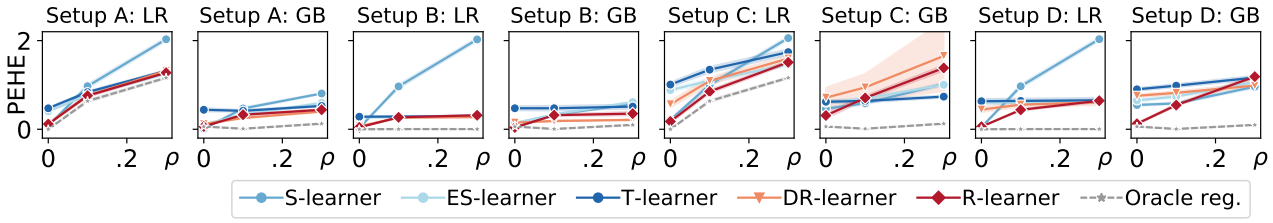


Figure 2. **Error in CATE estimation (PEHE) for the different candidate estimators in T .** All learners are implemented using linear regressions (LR) and extreme gradient boosting (GB), and considered across 4 different settings where the complexity of $\tau(x)$ increases in ρ. Shaded area indicates one SE.

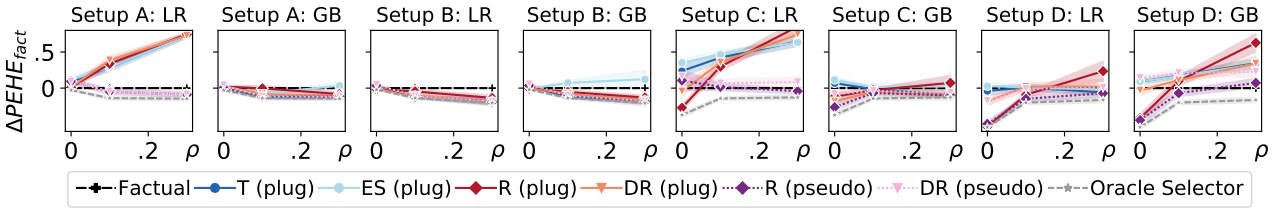


Figure 3. **Relative performance of different selection criteria.** Plotting PEHE_{fact}, the difference between the test-PEHE of the factual choice and the model selected by any given selection criterion (lower is better, negative means better than factual) implemented using linear regressions (LR) and extreme gradient boosting (GB). Each criterion gets access to T , i.e. the complete pool of 10 candidate estimators whose performance is shown in Fig. 2 above. Shaded area indicates one SE.

explore *the effects of model (mis)specification*. That is, whether algorithms (estimators and selection criteria) are given X or X as input X^{input} is the second experimental knob we consider: when given the original data X, this implicitly favors tree-based models like GB (this DGP mimics splits in decision trees) because a LR cannot fully recover X and hence not learn the patterns in either $\tau(\ )$ or $\mu_0(\ )$. Even when given the transformed data X as X^{input}, a simple LR cannot fit the POs $\mu_a(\ )$ due to the higher order interaction terms; however, the treatment effect itself is linear in X and could hence be fit with a LR if we observed $Y(1)$ $Y(0)$.

3. **Confounding.** Third, we compare confounded to randomized settings, where only in the former treatments are assigned based on variables that enter $\mu_0(\ )$. In the main text, the propensity score logits are linear in X^{input} and can hence be consistently estimated with logistic regressions in all settings. We consider additional settings with other propensities in Appendix D.4, where we also consider the effect of imbalance in treatment group sizes (whereas the main text has equal group sizes).

We vary ρ on the x-axis of our plots, which we split into four setups based on characteristics 2 and 3: Setup A is unconfounded and estimators & selectors get X as input, Setup B is unconfounded and estimators & selectors get X as input, Setup C is confounded and estimators & selectors get X as input and Setup D is confounded and estimators & selectors get X as input. A more formal description of the DGP can be found in Appendix C. Further, throughout we split training data of size n into $n_{train} = \frac{2n}{3}$ for training of all estimators and $n_{val} = \frac{n}{3}$ to be used by the selection criteria, and use an independent test-set of size $n_{test} = 500$ to evaluate a criterion by the test-set PEHE of its selected best model. We use $n = 1000 + 500$ as a default, but also consider the effect of having more ($n = 2000 + 1000$) or less ($n = 500 + 250$) data available in Appendix D.2. Every experiment is repeated for 20 random seeds, across which we report means and standard errors (SEs). Finally, we present additional results using the standard ACIC2016 and IHDP benchmarks in Appendix E, where we also outline why we believe they allow for less interesting analyses – highlighting further that careful design of DGPs was important to gain interesting insights.

5.2. Step 2: Examine performance of candidate estimators

Here, we briefly examine the performance of the underlying learners themselves in Fig. 2. We include an oracle that is trained directly on the (usually unknown) $\tau(X_i)$ to provide a lower bound on error due to misspecification (in particular, this highlights that in setup A and C, LRs cannot capture the CATE well while GBs could). We see that there are indeed interesting performance differences across learners and settings, meaning that no single learner, estimation strategy or ML method consistently dominates all others. As expected, R- and DR-learner show good performance especially when CATE is relatively simple. T- and ES-learners perform worst for $\rho = 0$, but relatively better for ρ large; the opposite is true for S-learners. Note that the performance of the LR S-learner is particularly poor for $\rho > 0$ because it can only learn a constant treatment effect (i.e. it is severely misspecified for $\rho > 0$). In Appendix D.4, we additionally consider imbalanced treatment group sizes and find that this worsens mainly the performance of the R-learner.

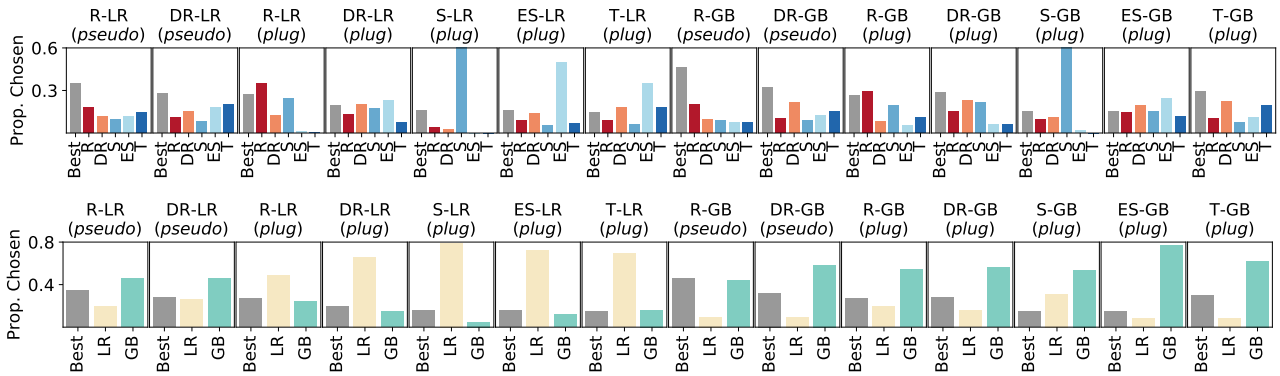


Figure 4. **Investigating the presence of congeniality bias between selection criteria and estimators: Estimation strategies (top) and ML methods (bottom).** Measuring the proportion of times the true best estimator is chosen (left-most bar in each plot), as well as what kind of estimator is chosen when the true best estimator is *not* chosen, across all settings considered.

5.3. Step 3: Towards understanding the performance of different selection criteria

We are now ready to examine the performance of the different model selection criteria: In Fig. 3 we present results of different criteria choosing between all 10 learner-method combinations. For legibility, we report performance in terms of ΔPEHE_{fact}, the difference between the test-PEHE of any given selector and the factual choice (lower is better, negative means better than factual).

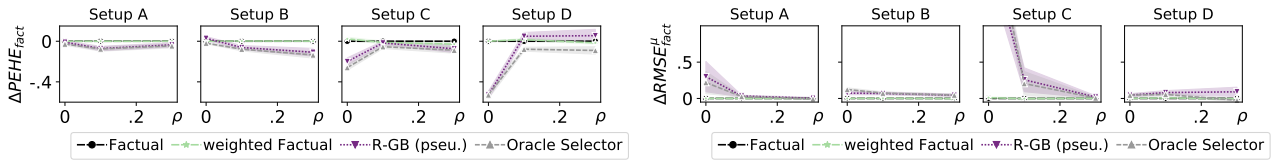
Q1: General performance trends. In Fig. 3, we observe that across all selectors, most performance differences and gains relative to factual performance are observed when i) treatment is randomized *and* the treatment effect is complex (ρ large in A&B), ii) there is confounding *and* the treatment effect is simple (ρ small in C&D) and iii) LRs are misspecified for the POs but not CATE (B& D). We observe that the performance of the plug-in criteria often follows the performance of their underlying methods (i.e. comparing the trends in Figs. 2 and 3). Further, the plug-in criteria based on ES- and T-learner generally perform the worst, especially when implemented using LRs. In setups C & D, the plug-in surrogate based on the R-learner works well when CATE complexity is low, but not when it is high – mimicking the pattern of the underlying estimator observed in the previous section. The pseudo-outcome criteria appear to perform best overall, with the R-pseudo-outcome often performing most similarly to the oracle selector. Note also that while the plug-in criteria based on indirect learners deteriorate substantially when misspecified (LR in Setups A & C), the pseudo-outcome criteria still perform well even when implemented using a misspecified model (LR).

Q2. Congeniality bias. Next we examine whether there is evidence for congeniality bias – i.e. whether plug-in or pseudo-outcome surrogate criteria appear to inherently favor estimators with similar inductive biases as the strategy used to provide a validation target. We propose to measure this by calculating the proportion of times an (i) estimation strategy or (ii) underlying ML method is selected by a validation criterion *whenever it does not identify the best estimator* (intuitively, we make this distinction because whenever an estimator is the oracle choice, selecting it should not be considered biased).[7] In Appendix D.3, we present a similar plot without making this distinction.

In Fig. 4 (top) we investigate congeniality between selection criteria and estimator strategy (i.e. R-, DR-, S-, ES- or T-learner, implemented using *either* ML method), pooled across all settings of Fig. 3. We observe that there is clear evidence for congeniality bias between some of the *plug-in* criteria and their corresponding learning strategy; this is most pronounced for the criteria relying on indirect learners, the plug-in S-learner and ES-learner in particular. The plug-in criterion based on the R-learner also clearly suffers from this, while the DR-plug-in criterion exhibits less of this behavior. The pseudo-outcome criteria overall display less pronounced preference for their own strategy, with the LR-implementations of pseudo-outcome R- and DR-criteria giving least evidence for such congeniality bias overall.

In Fig. 4 (bottom) we then investigate congeniality bias between selection criteria and estimator method (i.e. LR or GB, used with *any* estimation strategy). Also here we observe clear evidence for congeniality biases in almost all criteria: LR- (GB-)based criteria appear to prefer learners implemented using LR (GB). The only exception appears to be

[7] More formally, we investigate how often a selector E_j chooses the best estimator $\hat{\tau}$ that minimizes the true (unobservable) PEHE E_τ^{oracle} and, if not this top choice $\hat{\tau}$, which other type of estimator is chosen. That is, in Fig. 4, we measure using the leftmost bar labeled 'best' how often the correct best estimator is chosen –i.e. $\hat{P}(\arg\min_{\hat{\tau}_k} E_j(\hat{\tau}_k) = \hat{\tau}) = 1 \quad \alpha_{E_j}$, where α_{E_j} is the error-rate of the selector – , and other bars measure the proportion of times any specific type of estimator $\hat{\tau}_l \neq \hat{\tau}$ is chosen whenever E_j *does not make the right choice*. If some type of estimator is chosen disproportionately often – i.e. if $\hat{P}(\arg\min_{\hat{\tau}_k} E_j(\hat{\tau}_k) = \hat{\tau}_l | \hat{\tau}_l \neq \hat{\tau}) = 1 \quad \alpha_{E_j} >> \frac{\alpha_{\mathcal{E}_j}}{|T|}$ – we consider this evidence that the selector E_j may be biased towards choosing estimators of the type $\hat{\tau}_l$.



(a) CATE Estimation Performance

(b) Potential Outcome Estimation Performance

Figure 5. **Relative performance of different selection criteria when choosing between indirect learners only.** Left: Plotting PEHE_{fact}, the difference between the test-PEHE of the factual choice and the model selected by any given selection criterion. Right: Plotting RMSE^{μ}_{fact}, the difference between the average RMSE of estimating the potential outcomes using the factual choice and the model selected by any given selection criterion. (In both, lower is better and negative means better than factual.)

the LR-pseudo-outcome selectors, who actually select GB learners more often; this may partially explain their good relative performance compared to the other LR-based selectors.

Q3. What do we lose through factual evaluation? Finally, we consider the question of *why* we see in Fig. 3 that so many selectors using surrogates for the treatment effect can outperform factual selection in some scenarios. One possible explanation would be the exclusion of direct learners from the candidate pool available to E_Y^{fact}– however, as can be seen in Fig. 2, there is often *some* indirect learner that matches performance of a direct learner. Two alternative explanations we wish to test would be i) the presence of covariate shift due to confounding and ii) the incorrect focus of factual selection on performance in terms of estimating the POs. We can test i) by including an importance weighted factual selector, and ii) by restricting the candidate estimator pool available to *all selectors* to $T_{indirect}$ (i.e. excluding R- and DR-learner from the estimator pool).

In Fig. 5(left) we observe that i) does not seem to be the case as weighted and unweighted factual selection perform identically (one possible explanation for this is that none of the considered indirect estimators perform an internal covariate shift correction themselves). Considering ii) we do however observe that indeed both oracle and pseudo-outcome selectors appear to select different (better) indirect estimators than E_Y^{fact}, also in the absence of covariate shift (more saliently in Setup B). In Fig. 5(right), we show that these, in turn, indeed perform worse in terms of estimating the POs themselves, a trade-off that we expected in Sec. 4.

6. Conclusion

We studied the CATE model selection problem and focused on building understanding of the (dis)advantages of different model selection strategies – using factuals, plug-in surrogates or pseudo-outcome surrogates – that have been used or proposed in recent work. Instead of attempting to declare a global 'winner', we empirically investigated success- and failure modes of different strategies – and in doing so found that there *are* scenarios where factual selection can be appropriate but also scenarios where pseudo-outcome surrogate approaches are likely to perform better (only plug-in surrogate approaches seemed likely to underperform throughout). We hope that some of the insights presented here will give a starting point for practitioners able to identify how the likely characteristics of their own application translate to the scenarios we considered – for this purpose we include an additional digest of our findings in form of a Q&A with an imaginary reader in Appendix A. We also highlighted that there is a complex interplay between selection strategies, candidate estimators and the DGP used for testing – congeniality bias is likely to arise when the inductive biases of estimators and selection strategies align. By doing so, we also hope to have demonstrated to the community the need to conduct more simulation studies relying on carefully constructed DGPs to allow to disentangle different forces at play in this problem, enabling more nuanced analyses.

Limitations. Finally, note that we do not claim our results to be complete: to allow for interesting and nuanced insights, we needed to restrict our attention to specific questions and candidate estimators. We believe that there are a plethora of interesting questions to explore in this area, of which we only made an initial selection to serve as a starting point for discussion. It would, for example, be an interesting next step to consider how different criteria fare at selection between other classes of estimators, e.g. the method-specific neural-network-based estimators extending the work of Shalit et al. (2017), or, at a more microscopic level, at hyperparameter-tuning for any specific method. While our experimental results are limited to answering some of the questions we found most intriguing, we hope that the desiderata for experimental design that we discuss and implementations that we provide will allow future research to easily expand to further questions and associated DGPs.

Acknowledgements

We would like to thank Toon Vanderschueren, the members of the vanderschaar-lab and anonymous reviewers for insightful comments and discussions on earlier drafts of this paper. AC gratefully acknowledges funding from AstraZeneca.

References

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

Alaa, A. and Van Der Schaar, M. Validating causal inference models via influence functions. In *International Conference on Machine Learning*, pp. 191–201. PMLR, 2019.

Assaad, S., Zeng, S., Tao, C., Datta, S., Mehta, N., Henao, R., Li, F., and Duke, L. C. Counterfactual representation learning with balancing weights. In *International Conference on Artificial Intelligence and Statistics*, pp. 1972–1980. PMLR, 2021.

Athey, S. and Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Athey, S., Tibshirani, J., and Wager, S. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019.

Athey, S., Imbens, G. W., Metzger, J., and Munro, E. Using wasserstein generative adversarial networks for the design of monte carlo simulations. *Journal of Econometrics*, 2021.

Bica, I., Alaa, A. M., Lambert, C., and Van Der Schaar, M. From real-world patient data to individualized treatment effects using machine learning: current and future methods to address underlying challenges. *Clinical Pharmacology & Therapeutics*, 109(1):87–100, 2021.

Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., and Varoquaux, G. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*, pp. 108–122, 2013.

Chen, T. and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pp. 785–794, 2016.

Cui, Y. and Tchetgen, E. T. Selective machine learning of doubly robust functionals. *arXiv preprint arXiv:1911.02029*, 2019.

Curth, A. and van der Schaar, M. On inductive biases for heterogeneous treatment effect estimation. *Advances in Neural Information Processing Systems*, 34:15883–15894, 2021a.

Curth, A. and van der Schaar, M. Nonparametric estimation of heterogeneous treatment effects: From theory to learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pp. 1810–1818. PMLR, 2021b.

Curth, A., Svensson, D., Weatherall, J., and van der Schaar, M. Really doing great at estimating cate? a critical look at ml benchmarking practices in treatment effect estimation. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Dorie, V., Hill, J., Shalit, U., Scott, M., and Cervone, D. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019.

Hahn, P. R., Murray, J. S., Carvalho, C. M., et al. Bayesian regression tree models for causal inference: Regularization, confounding, and heterogeneous effects (with discussion). *Bayesian Analysis*, 15(3):965–1056, 2020.

Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., and Merrill, L. Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological bulletin*, 135(4):555, 2009.

Hassanpour, N. and Greiner, R. Counterfactual regression with importance sampling weights. In *IJCAI*, pp. 5880–5887, 2019a.

Hassanpour, N. and Greiner, R. Learning disentangled representations for counterfactual regression. In *International Conference on Learning Representations*, 2019b.

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

Holland, P. W. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

Johansson, F., Shalit, U., and Sontag, D. Learning representations for counterfactual inference. In *International conference on machine learning*, pp. 3020–3029. PMLR, 2016.

Johansson, F. D., Kallus, N., Shalit, U., and Sontag, D. Learning weighted representations for generalization across designs. *arXiv preprint arXiv:1802.08598*, 2018.

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

Knaus, M. C., Lechner, M., and Strittmatter, A. Machine learning estimation of heterogeneous causal effects: Empirical monte carlo evidence. *The Econometrics Journal*, 24(1):134–161, 2021.

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. Metalearners for estimating heterogeneous treatment effects using machine learning. *Proceedings of the national academy of sciences*, 116(10):4156–4165, 2019.

Louizos, C., Shalit, U., Mooij, J. M., Sontag, D., Zemel, R., and Welling, M. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.

Mahajan, D., Mitliagkas, I., Neal, B., and Syrgkanis, V. Empirical analysis of model selection for heterogenous causal effect estimation. *arXiv preprint arXiv:2211.01939*, 2022.

Nie, X. and Wager, S. Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika*, 108(2):299–319, 2021.

Parikh, H., Varjao, C., Xu, L., and Tchetgen, E. T. Validating causal inference methods. In *International Conference on Machine Learning*, pp. 17346–17358. PMLR, 2022.

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

Rosenbaum, P. R. and Rubin, D. B. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

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

Saito, Y. and Yasui, S. Counterfactual cross-validation: Stable model selection procedure for causal inference models. In *International Conference on Machine Learning*, pp. 8398–8407. PMLR, 2020.

Schuler, A., Jung, K., Tibshirani, R., Hastie, T., and Shah, N. Synth-validation: Selecting the best causal inference method for a given dataset. *arXiv preprint arXiv:1711.00083*, 2017.

Schuler, A., Baiocchi, M., Tibshirani, R., and Shah, N. A comparison of methods for model selection when estimating individual treatment effects. *arXiv preprint arXiv:1804.05146*, 2018.

Shalit, U., Johansson, F. D., and Sontag, D. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pp. 3076–3085. PMLR, 2017.

Wager, S. and Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018.

Wu, P. and Fukumizu, K. β-intact-vae: Identifying and estimating causal effects under limited overlap. *arXiv preprint arXiv:2110.05225*, 2021.

Yoon, J., Jordon, J., and Van Der Schaar, M. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.

Zhang, Y., Bellot, A., and van der Schaar, M. Learning overlapping representations for the estimation of individualized treatment effects. *arXiv preprint arXiv:2001.04754*, 2020.

Appendix

This Appendix is structured as follows: In Appendix A we provide an additional overview of key takeaways in form of a Q&A with an imaginary reader. In Appendix B we present an additional literature review discussing CATE estimators proposed in the recent ML literature. Appendix C discusses experimental details, and Appendix D presents additional results of the empirical study presented in the main text. Appendix E presents results on additional datasets.

A. Takeaways: A Q&A discussing the key insights from this paper with an imaginary reader

We studied the CATE model selection problem and focussed on building understanding of the (dis)advantages of different model selection strategies – using factuals, plug-in surrogates or pseudo-outcome surrogates – that have been used or proposed in recent work. Instead of attempting to declare a global 'winner', we empirically investigated success- and failure modes of different strategies in the hope that some of the insights presented here will give a starting point for practitioners able to identify how the likely characteristics of their own application translate to the scenarios we considered. For this purpose and because our results may be extensive to digest, we present a light discussion of takeaways in form of a Q&A with an imaginary reader below.

Disclaimer: as we emphasized above, our results do not and cannot cover all possible scenarios. We answer the questions below based on our own empirical studies as well as intuition we built throughout the paper, but note that these conclusions are based only on the limited (yet nuanced) scenarios we were able to consider.

- *Q: So what is the best model selection criterion?* A: There are no magic bullets (yet?!).
- *Q: Fine. What are good candidates then?* A: This appears to depend on your data, but overall we found pseudo-outcome surrogates and factual selection to perform well in different scenarios.
- *Q: Let me tell you something about my data then. I have confounded data. What should I do?* A: We observed in experiments that especially when the treatment effect is simple and data is confounded, pseudo-outcome criteria using the R- or DR-objective perform much better than other criteria. When the treatment effect is a more complex function, factual criteria appear to perform better.
- *Q: What if I have unconfounded data from a clinical trial?* A: When data is unconfounded and models are correctly specified, we found that the selection criterion has slightly less influence. Only in the setting (B) where the treatment effect is a much simpler function than the POs we observed some improvements in using other criteria – in this case both plug-in and pseudo-outcome surrogates performed better as ρ increased.
- *Q: I expect the treatment effect heterogeneity to be relatively less pronounced than heterogeneity in outcomes regardless of treatment (there is much more prognostic rather than predictive information). What should I do?* A: Pseudo-outcome surrogates, especially the R-learner objective, appear to work very well in this scenario.
- *Q: I expect the opposite – treatment effects are likely a very complex function of characteristics – what does that mean for me?* In this case, using any surrogates may introduce more noise than they help (especially when datasets are small and confounded); it may be advisable to rely on factual validation in this case.
- *Q: I would prefer to rely on factuals for validation because I trust that the most as it doesn't require me to estimate any additional parameters. What do I lose out on?* A: We observed that using factual evaluation can be worse in some scenarios *both* because it means that you can only evaluate a smaller set of estimators *and* because it targets the wrong objective – this particularly seems to matter when the POs are very complex relative to CATE.
- *Q: What is this "congeniality bias" you were referring to?* A: We consider congeniality bias the issue that surrogate validation targets may advantage CATE estimators $\hat{\tau}_k(x)$ that are *structurally similar* to the used surrogate, due to being imbued with similar inductive biases. The term itself is usually used in the psychology literature to indicate that individuals may have a systematic preference for information consistent with current beliefs (Hart et al., 2009).
- *Q: I see! Which type of selection criteria suffer from this the most?* A: In our experiments we found that plug-in surrogate selection criteria appear to suffer from this more than pseudo-outcome selection criteria.
- *Q: This is all very insightful – but you do not cover an aspect of model selection I would be interested in. Do you have any advice for me to design my own empirical study investigating this question?* We agree that there are many more interesting questions and hope to expand on these in the future! In Section 4.2 we present desiderata for insightful experiment design if you want to design your own – and we will also release our code in the future, which is set up in such a way to allow plug-and-play with new selection criteria and datasets!

B. Additional Literature Review

B.1. Further CATE estimation strategies.

In the main text we focussed on one of two prominent streams in the CATE estimation literature, namely the one relying on so-called meta-learners that can be easily implemented using *any ML prediction method*, as originally proposed by Künzel et al. (2019) and later extended by Nie & Wager (2021); Kennedy (2020); Curth & van der Schaar (2021b).

Many methods proposed recently within the ML literature, however, belong to a second stream of literature focussed on *adapting specific* ML methods to the CATE estimation context. This literature has overwhelmingly relied on indirect estimation strategies, proposing models that learn to make unbiased predictions of outcome under each treatment (so that CATE can be estimated as thei difference). Very popular in this literature, as originally proposed by Johansson et al. (2016); Shalit et al. (2017) and later extended in e.g. Johansson et al. (2018); Hassanpour & Greiner (2019a;b); Assaad et al. (2021); Curth & van der Schaar (2021a;b) has been the use of neural networks to learn a shared representation $\Phi(x)$ that is then used by two treatment-specific output heads h_a to estimate $\hat{\mu}_a(x) = h_a(\Phi(x))$. In this framework, covariate shift arising due to confounding is usually addressed by addition of a balancing regularization term penalising the discrepancy in distribution of $\Phi(x)$ between treatment groups (Shalit et al., 2017) and importance weighting (Johansson et al., 2018; Hassanpour & Greiner, 2019a; Assaad et al., 2021). Other work has investigated the use of Gaussian Processes (Alaa & van der Schaar, 2018), GANs (Yoon et al., 2018), VAEs (Louizos et al., 2017; Wu & Fukumizu, 2021) and deep kernel learning (Zhang et al., 2020) for PO estimation.

The statistics and econometrics literatures, on the other hand, next to the meta-learner strategies discussed in the main text, have mainly relied on tree-based methods, most prominently using S-learner style BART (Hill, 2011) and direct estimators in the form of causal trees (Athey & Imbens, 2016) and causal forests (Wager & Athey, 2018; Athey et al., 2019; Hahn et al., 2020).

B.2. Further model selection strategies.

The model selection strategies discussed in the main text, i.e. those proposed and/or studied in Rolling & Yang (2014); Nie & Wager (2021); Saito & Yasui (2020); Alaa & Van Der Schaar (2019); Schuler et al. (2018); Mahajan et al. (2022) all use the given observational data to estimate model performance directly, imputing only some nuisance parameters or surrogate targets. A separate strand of literature (Schuler et al., 2017; Athey et al., 2021; Parikh et al., 2022), which we did not consider further here, suggests to instead validate causal inference models by learning a generative model from the observational dataset at hand and use it to simulate multiple test datasets that share some characteristics with the dataset of interest but have *known* treatment effect that can be used to compare treatment effect estimates of candidate estimators. The estimator found to perform best on such generated datasets should then be used on the real data (Schuler et al., 2017; Athey et al., 2021; Parikh et al., 2022). Finally, other interesting model selection problems exist in the treatment effect estimation context, e.g. the question of how to choose the best first stage *nuisance estimators* for multi-stage treatment effect estimators considered in Cui & Tchetgen (2019).

C. Experimental Details

C.1. Implementation details

All code is written in python in sklearn-style to allow for modularity and ease of reuse. All code is available at `https://github.com/AliciaCurth/CATESelection` as well as the vanderschaar-lab repository `https://github.com/vanderschaarlab`.

Underlying ML methods

- For linear regressions (LR) with ridge (l2) penalty, we use `RidgeCV` as implemented in `sklearn` (Buitinck et al., 2013); this allows us to automatically implement a sweep over ridge penalties $\lambda \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3, 10^4\}$ whenever LR is used as a subroutine in any meta-learner (to estimate $\mu(x)$, $\mu_0(x)$, $\mu_1(x)$, or $\tau(x)$).
- For extreme gradient boosted trees (GB), we use the `XGBRegressor` as implemented in the `xgboost` python package (Chen & Guestrin, 2016). Whenever GB is used as a subroutine in any meta-learner, we use `sklearn`'s 5-fold `GridSearchCV` to perform a sweep over all combinations of `learning_rate` $\in \{.1, .3\}$, `max_depth`

$2\ f1,3,6g$ and `n_estimators` $2\ f20,100g$, which are hyperparameters that we observed to have an impact on all learners.

- Finally, for logistic regressions that used to estimate propensity scores $\pi(x)$ we use the `sklearn` implementation `LogisticRegressionCV`, allowing us to sweep over regularization parameters $\lambda\ 2\ f10^{\ 5}, 10^{\ 3}, 10^{\ 2}, 10^{\ 1}, 1g$ whenever it is used.

Note that we reoptimize hyperparameters as part of all subroutines M, which means they are chosen anew for every meta-learner in every seed of every experiment.

Meta-learners Given a ML-method M, implemented as discussed above, we train the different meta-learners as follows:

- S-learner: Append A to X to give $X^0 = (X, A)$. Call M`.fit`(X^0, Y).
- Extended S-learner (ES-learner): Append A and $A \quad X$ to X to give $X^0 = (X, A \quad X, A)$. Call M`.fit`(X^0, Y).
- T-learner: Separate data by treatment indicator. Call M`.fit`$(X[A == a], Y[A == a])$ for each treatment group $a\ 2\ f0,1g$
- DR-learner: Call T-learner to get estimates of $\mu_a(x)$. Fit propensity estimate by calling M`.fit`(X, A). Use these estimates to compute pseudo-outcome $Y_{DR,\eta}$ as specified in the main text. Call M`.fit`$(X, Y_{DR,\eta})$. (Note: we also tested using 5-fold cross-fitting as suggested by Kennedy (2020) to ensure consistency, but did not find this to improve performance).
- R-learner: Fit propensity estimate as in DR-learner. Fit unconditional mean estimate $\mu(x)$ by calling M`.fit`(X, A). Compute pseudo-outcome $Y_{R,\eta} = \frac{Y_i \quad \mu(X_i)}{(A_i \quad \pi(X_i)}$ and weights $\beta_i = (A_i \quad \tilde{\pi}(X_i))^2$. Call M`.fit`$(X, Y_{R,\eta},$ `sample_weight` $= \beta)$. (Note: we also tested using 5-fold cross-fitting as suggested by Nie & Wager (2021) to ensure consistency, but did not find this to improve performance).

Selection criteria All selection criteria are computed by solely considering validation data. For plug-in surrogate criteria, the strategies discussed above are used on the validation data to compute a plug-in estimate of the treatment effect.

For the pseudo-outcome surrogate criteria, we perform 5-fold cross-fitting to avoid correlation between nuisance estimates and outcomes; that is we split the validation data into 5 folds, and use only the 4 folds a data-point is not in to impute their pseudo-outcome.

While not presented in the main text, in Appendix D.1 we also computed pseudo-outcome surrogates using the PW-pseudo outcome $Y_{PW,\eta} = \left(\frac{A}{\hat{\pi}(X)} \quad \frac{(1 \quad A)}{1 \quad \hat{\pi}(X)}\right) Y$ and the matching pseudo outcome of Rolling & Yang (2014), computed by finding the nearest neighbor in Euclidean distance. We also computed the Alaa & Van Der Schaar (2019)'s influence function validation criterion, we amounts to selecting $\hat{\tau}_k(x)$ that minimizes

$$Y_{IF} = (1 \quad B)\tilde{\tau}(x)^2 + BY(\tilde{\tau}(x) \quad \hat{\tau}(x)) \quad D \quad (\tilde{\tau}(x) \quad \hat{\tau}_k(x))^2 + \hat{\tau}_k(x)^2 \tag{2}$$

with $D = A \quad \tilde{\pi}(x)$, $B = 2AC^{\ 1}$ and $C = \tilde{\pi}(1 \quad \tilde{\pi})$ and $\tilde{\tau}(x)$ is a T-learner estimate. All nuisance parameters are estimated using 5 fold cross-estimation.

C.2. Data-generating process (DGP)

We build on the DGP used in (Curth & van der Schaar, 2021a) for our experiments. The main differences lie in that we a) randomly binarize the data to consider the effect of misspecification, b) consider higher order interactions to make differences between CATE and the POs more salient and c) also induce confounding.

We also use the covariate data from the Collaborative Perinatal Project provided[8] for the first Atlantic Causal Inference Competition (ACIC2016) (Dorie et al., 2019) and process all covariates according to the transformations used for the competition[9]. The original dataset has $d = 58$ covariates, of which we exclude the 3 categorical ones. Of the remaining 55 covariates, 5 are binary, 27 are count data and 23 are continuous. Because we found that the existing binary and count data are very sparse, we instead decided to randomly binarize variables, by choosing a random observed value in each column and keep only the 23 continuous columns to resulting in a new input dataset X used to create a DGP that mimics a decision tree. (Note that, while not used for outcome simulation, all other columns remain part of X^{input} given to estimators and selection criteria, so all estimators have to also learn to distinguish informative from uninformative columns.)

[8] This can be retrieved from `https://jenniferhill7.wixsite.com/acic-2016/competition`

[9] We use the code at `https://github.com/vdorie/aciccomp/blob/master/2016/R/transformInput.R`

Similar to Curth & van der Schaar (2021a), we then use the input data in a linear model with interaction terms:

$$Y_i = c + \sum_{j=1}^{d} \beta_j X_j + \sum_{j,l}^{d} \beta_{j,l} X_j X_l + \sum_{j,l,k}^{d} \beta_{j,l,k} X_j X_l X_k + \sum_{j,l,k}^{d} \beta_{j,l,k,m} X_j X_l X_k X_m + A_i \sum_{j=1}^{d} \gamma_j X_j + \epsilon_i \quad (3)$$

where ϵ_i $N(0, .1)$, β_j $B(.3)$ and γ_j $B(\rho)$. We include each variable randomly into one first, second and third-order interaction term, for which we then simulate coefficient β $B(.2)$. We chose for each coefficient to be binary to avoid large variances in the scale of POs and CATE across different runs of a simulation, such that RMSE remains comparable across runs.

We then simulate confounding by assigning treatment according to a propensity score $\pi(x) = expit(\xi Z(X^{input}\beta))$ where $Z(\,)$ denotes standardisation across the simulation data and β is the linear coefficient from eq. 3, ensuring that all variables are true confounders. Note also that X^{input} –i.e. the data as observed by estimators and selectors – enters the propensity score, ensuring that a logistic regression is always correctly specified for estimating the propensity score. We experiment with further settings in Appendix D.4.

The three main experimental knobs under consideration are thus CATE complexity ρ, confounding strength ξ and estimator & selector access to input data X versus X . In our experiments we always vary ρ 2 f0, .1, .3g and define settings:

- A: Continuous input data $X^{input} = X$, no confounding $\xi = 0$
- B: Binarized input data $X^{input} = X$, no confounding $\xi = 0$
- C: Continuous input data $X^{input} = X$, no confounding $\xi = 3$
- D: Binarized input data $X^{input} = X$, no confounding $\xi = 3$

D. Additional Results Using the Main DGP

D.1. Additional selection criteria

In Fig. 6 we additionally show performance of matching, two further pseudo-outcomes (PW- and RA-pseudo-outcomes considered in Curth & van der Schaar (2021b)), influence function (IFs) and weighted factual validation (wFactual) not presented in the main text. We excluded them in the main text for legibility and because they did not present any improvements over factual selection (if anything they usually performed worse).

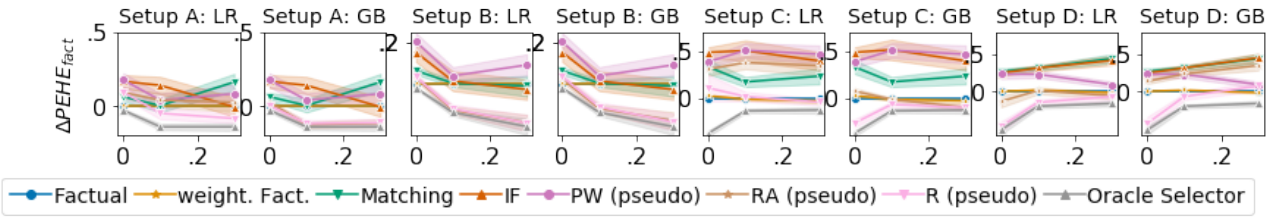
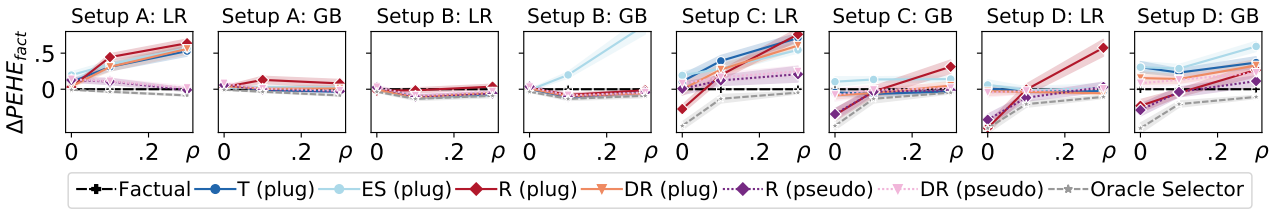
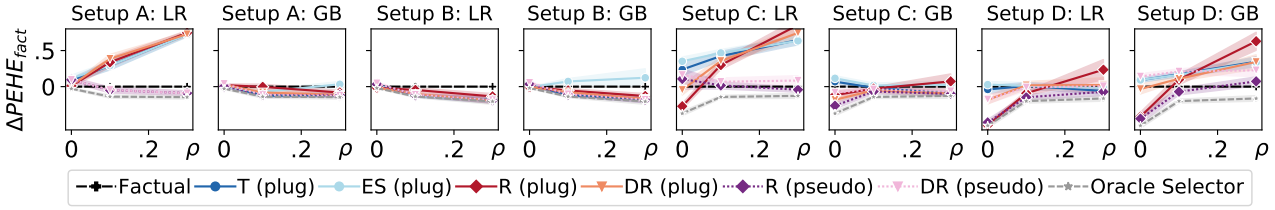


Figure 6. PEHE$_{fact}$, the difference between the test-PEHE of any given selection criterion and the factual choice (lower is better, negative means better than factual), for different selection criteria, implemented using linear regressions (LR) and extreme gradient boosting (GB) across 4 different settings, including additional criteria: RA- & PW-pseudo-outcomes, matching, influence function (IFs) and weighted factual validation (wFactual) not presented in the main text. Here, the complexity of $\tau(x)$ increases in ρ. Shaded area indicates one SE.

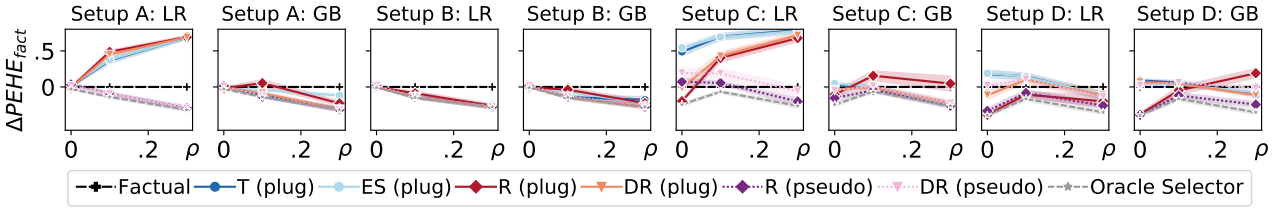
D.2. Additional sample sizes



(a) n = 500 + 250



(b) n = 1000 + 5000 (reproduced from main text)



(c) n = 2000 + 1000

Figure 7. **Relative performance of different selection criteria at multiple sample sizes.** Plotting PEHE_{fact}, the difference between the test-PEHE of the factual choice and the model selected by any given selection criterion (lower is better, negative means better than factual) implemented using linear regressions (LR) and extreme gradient boosting (GB).

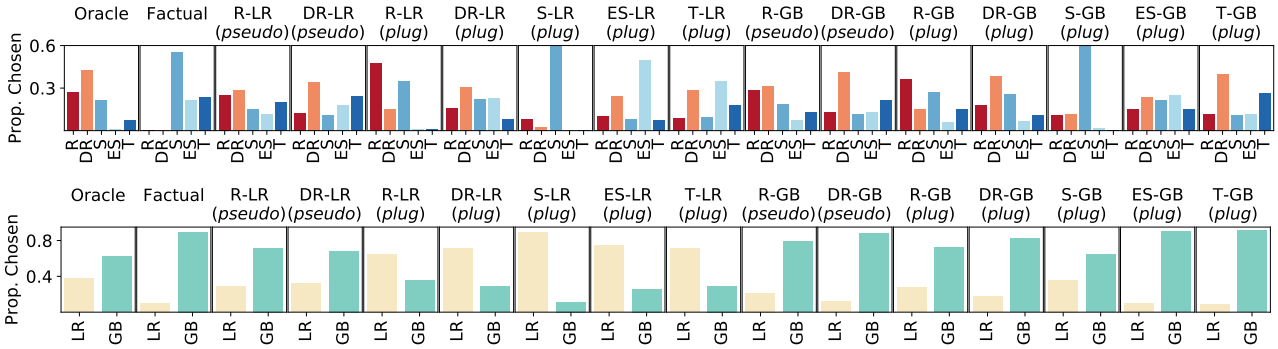


Figure 8. **Investigating the presence of congeniality bias between selection criteria and estimators: Estimation strategies (top) and ML methods (bottom).** Measuring the proportion of times a specific estimator is chosen, across all settings considered in the main text. Note that, different from Fig. 4, we measure the *absolute* proportion of time any estimator is chosen in this Figure (instead of how often an estimator-type is chosen when a mistake is made as in the main text).

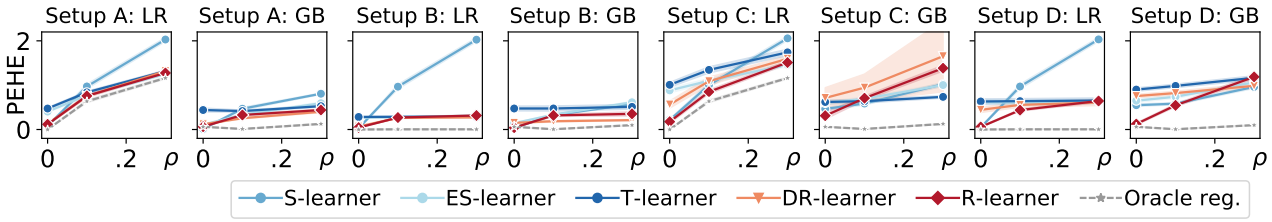
D.3. Additional congeniality plots

In Fig. 8, we present the *absolute* number of times any estimator type is chosen by any selection criterion – this is different from from Fig. 4 in the main text, which focussed on the types of estimators that are chosen when a selector *makes an error*. While congeniality is much more obvious when we consider only the errors made by selectors, some of the congeniality

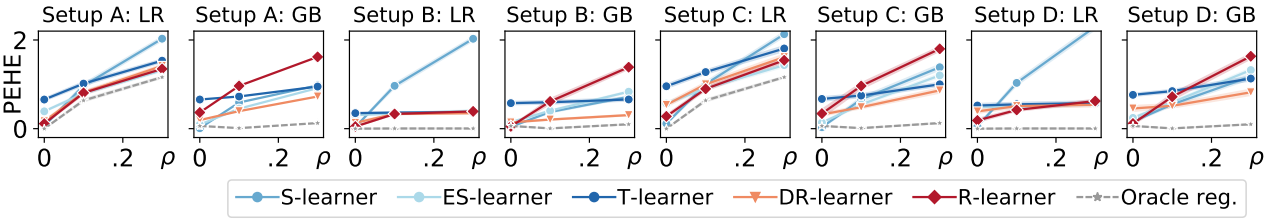
patterns disucssed in the main text are also clearly reflected in Fig. 8.

D.4. Additional settings with other propensities

Imbalance. In Figs. 9 and 10 we investigate the effects of adding imbalanced treatment group sizes by re-scaling treatment assignment propensities. Instead of balanced marginal treatment propensity $\pi = 0.5$ considered in the main text, we now assign treatment with marginal propensity $\pi = .2$ so that there are substantially more control than treatment units. In Fig. 9, in terms of underlying learners, we observe that the most salient difference is that the R-learner now performs relatively worse at large effect heterogeneity. In Fig. 10, we find that in the imbalanced setting there is much less improvement over factual selection in the confounded settings with small ρ, which is where in the balanced settings there were most performance gains.



(a) π = 0.5 (reproduced from main text)

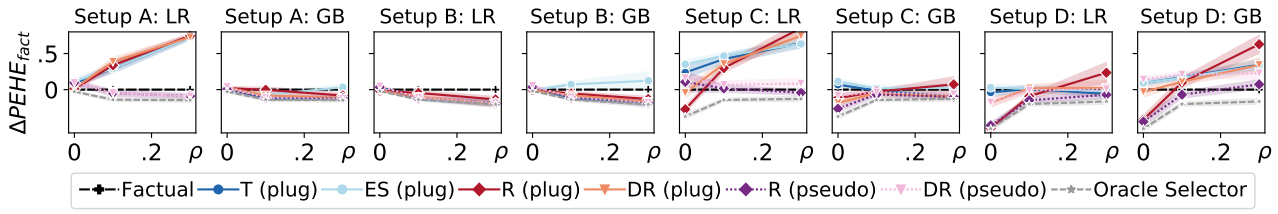


(b) π = 0.2

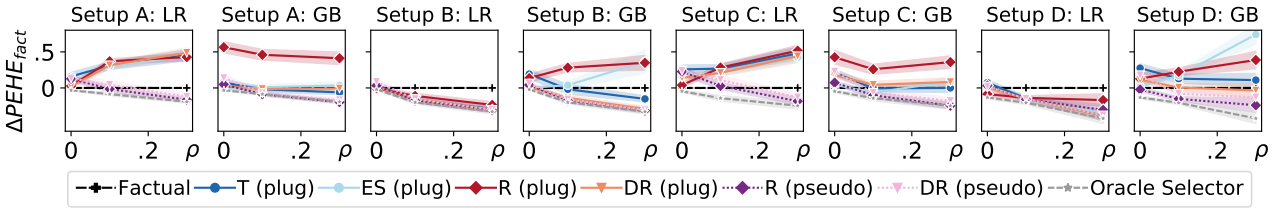
Figure 9. **Error in CATE estimation (PEHE) for the different candidate estimators in T, without treatment-group imbalance (top) and with imbalance (bottom).** All learners are implemented using linear regressions (LR) and extreme gradient boosting (GB), and considered across 4 different settings where the complexity of $\tau(x)$ increases in ρ. Shaded area indicates one SE.

Other propensity specifications. In Figs. 11 and 12, we investigate further variation on the used propensity score specification. In particular, note that as discussed in Appendix C, we ensured that the propensity score is always correctly specified using a logistic regression on X^{input}. As a by-product, this means that in Setup C considered in the main text, π is a function of X and μ is a function of X , while in Setup D, both π and μ are functions of the binarized (observed) X and hence more aligned. Here, we therefore consider two additional setups where π is a function of the version of the covariates that is not observed (i.e. X for Setup C*, where estimators are given X, and X for Setup D*, where estimators are given X). This could be expected to affect the results in two ways: on the one hand, in the new setup C* and D*, propensity score estimators are misspecified, which could negatively affect estimators and selectors relying on those. On the other hand, note that in the old setup D and the new setup C*, π and μ depend on the same transformation of the covariates X , while in the old setup C and the new setup D*, π and μ do not depend on the same transformation of the covariates – in D and C*, propensity scores and outcomes are more aligned, which generally makes estimation harder.

In Figs. 11 and 12, we observe that the second effect appears to outweigh the first: especially when using misspecified models (LRs in setups C and C*), for both the candidate estimators and the selection criteria, we observe that aligning π and μ more (i.e. moving from setup C to C*) deteriorates their performance – regardless of whether a propensity score actually needs to be estimated (e.g. the performance of S- and T-learners, which do not include propensity estimates, also deteriorates).



(a) $\pi = 0.5$ (reproduced from main text)



(b) $\pi = 0.2$

Figure 10. **Relative performance of different selection criteria without treatment-group imbalance (top) and with imbalance (bottom).** Plotting PEHE_{fact}, the difference between the test-PEHE of the factual choice and the model selected by any given selection criterion (lower is better, negative means better than factual) implemented using linear regressions (LR) and extreme gradient boosting (GB). Each criterion gets access to $\mathcal{T}$, i.e. the complete pool of 10 candidate estimators whose performance is shown in Fig. 9 above. We consider 4 different settings, where the complexity of $\tau(x)$ increases in ρ. Shaded area indicates one SE.

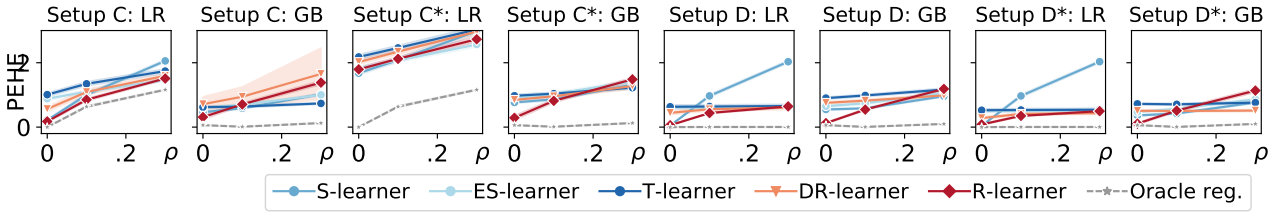


Figure 11. **Error in CATE estimation (PEHE) for the different candidate estimators in $\mathcal{T}$, for confounded settings with different propensity score specifications.**

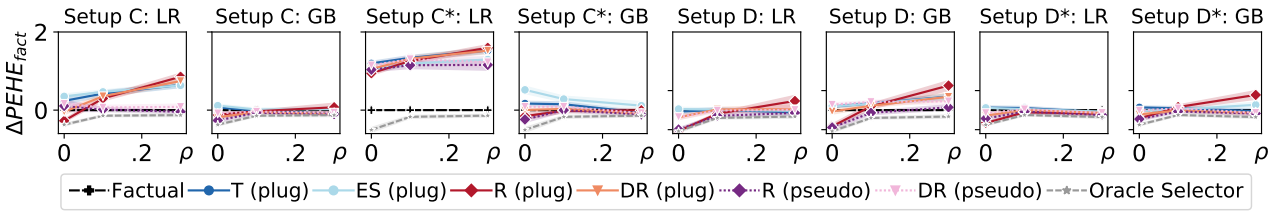


Figure 12. **Relative performance of different selection criteria, for confounded settings with different propensity score specifications.**

E. Additional Datasets: IHDP and ACIC2016

For completeness, we repeat the experiments conducted in Curth et al. (2021) who highlighted some problems with the use of the ACIC2016 and IHDP datasets in the current CATE estimation literature, in particular the lack of regard for the underlying structural characteristics of these datasets (the setting of experimental knobs they are 'biased' towards). With this in mind, we repeat the same experiments from the main text using the setups discussed in Curth et al. (2021); for details refer to their paper.

In particular, note that the original IHDP dataset as constructed by Hill (2011) has a very complex treatment effect function,

likely advantaging indirect learners. Curth et al. (2021) construct a modified version where the POs remain complex but the treatment effect is simply linear, now possibly advantaging direct learners. Additionally, Curth et al. (2021) study 3 (numbers 2, 26 and 7) of the 77 settings of the ACIC2016 competition (Dorie et al., 2019): these differ only in the 'CATE complexity', where setting 2 has no heterogeneity, 26 has some and 7 is very heterogeneous – setting similar expectations for relative performance. The IHDP dataset is very small already, with a total size of 747 so we retain the original train test split and split the training data $2/3$ to $1/3$ for training and validation, as before. The ACIC2016 dataset use the same underlying covariates as our simulations, thus we consider the same three sample size settings. Finally, note that Curth et al. (2021) observed very high variance in absolute RMSE across settings even when using the same estimation strategy. To stabilize results we therefore report relative RMSE ($RMSE(method)/RMSE(baseline)$) where the baseline for all selection criteria remains factual selection, and the baseline for learners on IHDP is T-LR and T-GB on ACIC2016 (the respective best performing T-learners).

Comparing underlying learner performance. We report relative performance of the underlying learners in Table E. We observe that the relative performance of underlying learners is as expected on both datasets and mimics what was shown in Curth et al. (2021): direct learners have an advantage on the DGPs with simpler CATE, while T-learners have an advantage on the settings with complex CATE. We also observe that the R-GB learner generally performs worse than in the main text even at low CATE complexity, which may be due to the imbalance in treatment group sizes in all datasets.

Table 1. Relative PEHE of underlying learners on the IHDP and ACIC settings. Averaged across all 100 simulations for IHDP, and across 10 each for ACIC.

Setting	Oracle-GB	Oracle-LR	S-GB	S-LR	ES-GB	ES-LR	T-GB	T-LR	DR-GB	DR-LR	R-GB	R-LR
Original IHDP	0.58	0.78	1.58	2.58	1.54	1.58	1.14	1.00	1.25	1.11	2.60	1.65
Modified IHDP	0.05	0.00	0.85	0.46	1.07	1.24	1.35	1.00	0.94	0.67	4.46	1.54
ACIC 2, n=750	0.00	0.00	0.31	0.32	0.52	0.92	1.00	1.10	0.64	0.46	1.06	0.31
ACIC 2, n=1500	0.00	0.00	0.37	0.15	0.53	0.84	1.00	1.08	0.67	0.40	0.91	0.29
ACIC 2, n=3000	0.00	0.00	0.50	0.20	0.69	0.93	1.00	1.04	0.49	0.47	1.33	0.31
ACIC 26, n=750	0.59	1.07	1.12	1.61	1.04	1.26	1.00	1.28	0.94	1.26	1.24	1.27
ACIC 26, n=1500	0.58	1.32	1.20	2.03	1.08	1.57	1.00	1.47	0.91	1.46	1.22	1.47
ACIC 26, n=3000	0.58	1.55	1.20	2.41	1.03	1.74	1.00	1.71	0.95	1.70	1.40	1.72
ACIC 7, n=750	1.10	1.27	1.15	1.71	1.08	1.47	1.00	1.44	0.95	1.46	1.37	1.44
ACIC 7, n=1500	0.68	1.56	1.23	2.20	1.10	1.75	1.00	1.71	1.01	1.77	1.43	1.74
ACIC 7, n=3000	0.64	1.74	1.25	2.54	1.09	1.93	1.00	1.90	0.99	1.91	1.57	1.91

Comparing selector performance. We report relative performance of selectors in Table E. Relative selector performance on IHDP is largely as expected, except that pseudo R- and DR-criterion perform worse than expected on the modified setting, while the plug-in criteria perform better. It is difficult to pinpoint an origin for this, because the IHDP dataset also has i) a much larger control than treated population, ii) limited overlap and iii) very small sample size. Results on the ACIC datasets are also mixed; here we observe improvements over factual selection mainly for the smallest datasets and when CATE is simple (setting 2). It is possible that this is partially due to the fact that *no* method is able to fit the DGP particularly well with high heterogeneity, seeing as oracle selector performance is only marginally better than factual selection in settings 26 and 7. Note that these ACIC simulations also have limited overlap and imbalances between treatment and control group. We hope that this discussion highlights why we deemed it necessary to construct our own DGPs: Because all these forces are deeply entangled in existing datasets, it is extremely difficult to use them to disambiguate the effects of different factors on performance.

Table 2. PEHE of model selection metrics relative to factual selection on the IHDP and ACIC settings. Averaged across all 100 simulations for IHDP, and across 10 each for ACIC.

Setting	Oracle	Factual	S-GB	S-LR	ES-GB	ES-LR	T-GB	T-LR	PlugDR GB	PlugDR LR	PlugR GB	PlugR LR	PseuDR GB	PseuDR LR	PseuR GB	PseuR LR
Original IHDP	0.91	1.00	1.53	2.57	1.53	1.43	1.02	1.02	1.12	1.10	1.81	1.63	1.03	1.03	1.10	1.11
Modified IHDP	0.58	1.00	0.91	0.71	0.97	1.36	0.85	0.79	0.77	0.76	3.32	1.91	1.07	1.01	1.36	1.38
ACIC 2, n=750	0.45	1.00	2.38	6.86	2.38	5.10	3.11	3.99	3.20	0.78	5.57	2.45	3.50	5.80	4.50	4.56
ACIC 2, n=1500	0.35	1.00	0.54	1.13	0.98	0.90	2.27	1.83	2.08	1.29	2.31	0.51	2.70	2.80	2.16	2.74
ACIC 2, n=3000	0.35	1.00	0.62	0.39	1.12	1.57	4.16	1.79	2.96	1.04	3.42	1.55	4.83	4.94	4.48	4.21
ACIC 26, n=750	0.91	1.00	1.20	1.65	1.12	1.27	0.95	1.23	0.97	1.26	1.11	1.31	1.00	0.96	0.95	0.96
ACIC 26, n=1500	0.88	1.00	1.25	2.00	1.16	1.48	0.90	1.41	0.93	1.46	1.30	1.46	0.98	1.00	0.94	1.00
ACIC 26, n=3000	0.86	1.00	1.25	2.46	1.08	1.79	1.00	1.78	1.01	1.78	1.27	1.77	1.06	1.08	1.12	1.09
ACIC 7, n=750	0.85	1.00	1.34	1.65	1.12	1.38	0.90	1.35	0.90	1.32	1.21	1.46	0.90	0.90	0.88	0.91
ACIC 7, n=1500	0.94	1.00	1.24	2.13	1.20	1.72	1.00	1.66	1.03	1.66	1.35	1.78	0.99	1.04	1.08	1.07
ACIC 7, n=3000	0.98	1.00	1.23	2.53	1.12	1.91	0.99	1.89	1.05	1.92	1.25	1.99	0.98	1.04	1.08	1.04