A Model Confidence Set for Learner Selection in Generic Machine Learning (DRAFT)

Pierce Plucker*

March 27, 2026

Abstract

Generic Machine Learning (Chernozhukov et al., 2025a) provides valid inference on heterogeneous treatment effects using any machine learning algorithm to proxy the conditional average treatment effect, but offers limited guidance on which learner's results to report when multiple learners are employed. I propose applying the model confidence set of Hansen et al. (2011) to the split-level goodness-of-fit values produced by the GML procedure, yielding a data-driven set of learners whose performances are statistically indistinguishable. Inference then proceeds over this set rather than the full collection of candidates, with overall coverage guaranteed at $1 - \alpha_1 - \alpha_2$ by a budget-splitting argument. Monte Carlo simulations show that the procedure controls size, sacrifices little power relative to naive selection, and exposes meaningful variation in point estimates across equally credible learners that is invisible under standard reporting practices.

1 Introduction

Generic Machine Learning (Chernozhukov et al., 2025a) provides a powerful framework for inference on heterogeneous treatment effects in randomized experiments. The method is deliberately agnostic with respect to the machine learning algorithm used to proxy the conditional average treatment effect (CATE): any learner—or collection of learners—may be employed in the first stage, and valid inference on key parameters follows from sample splitting and median aggregation in the second stage. This agnosticism is a central feature of the framework, as researchers typically have little basis for specifying which functional form best approximates the true CATE *a priori*.

In practice, however, most applied work using GML has relied on a small, recurring set of four to five learners, commonly random forests, gradient boosting, regularized regression, and neural

*Department of Economics, Oklahoma State University. Email: pierce.plucker@okstate.edu

networks. This convention is understandable: with a small learner set, one can apply a Bonferroni correction across all learners and still obtain reasonably tight confidence intervals. But this approach is at odds with the agnostic philosophy of GML. A richer set of candidate learners increases the probability that at least one learner captures the relevant structure in the true CATE, particularly when the functional form is complex or involves interactions among covariates. The cost of expanding the learner set under current practice is a loss of statistical power, since a Bonferroni correction over L learners scales the critical value by L. As L grows, the correction becomes increasingly conservative and the inferential gains from a better proxy may be offset by wider confidence intervals.

This tension between exploration and power gives rise to a learner selection problem. If the researcher could identify a small subset of learners whose performance is statistically indistinguishable from the best, she could apply a multiplicity correction only over this subset, preserving power while maintaining the benefits of a large initial candidate set. The challenge is to construct such a subset in a principled, data-driven manner. Chernozhukov et al. (2025a) suggest reporting the union of confidence intervals or applying a Bonferroni correction when two learners are "near-winners," but leave open the question of how to determine which learners qualify as near-winners and how to proceed when more than two are competitive.

This paper proposes a solution based on the model confidence set (MCS) of Hansen et al. (2011). The MCS is a sequential testing procedure that begins with the full set of candidate learners and eliminates those whose goodness-of-fit is significantly inferior, terminating when the surviving learners cannot be statistically distinguished from one another. The resulting set contains the best learner with a user-specified probability, analogous to a confidence interval for a parameter. When applied to GML, the MCS provides a formal answer to the near-winner question: the near-winner set is precisely the model confidence set, and its cardinality adapts to the informativeness of the data. In settings where a single learner clearly dominates, the MCS shrinks to a singleton and no multiplicity correction is needed. When several learners perform comparably, the set expands and the correction is applied only over the relevant competitors.

I establish the theoretical validity of this two-stage procedure–MCS for learner selection followed by union-based inference over the selected set–and show that the overall coverage probability is bounded below by $1 - \alpha_1 - \alpha_2$, where α_1 governs the MCS and α_2 governs the inference stage. Monte Carlo simulations using a benchmark DGP from the recent discussion of GML (Wager, 2025; Chernozhukov et al., 2025b) demonstrate that the procedure controls size, preserves power relative to naive selection, and provides substantive information about which learners are credible proxies for the CATE. Beyond hypothesis testing, the MCS offers a practical diagnostic: the range of point estimates across the surviving learners quantifies the sensitivity of the reported magnitude to the choice of learner, information that is unavailable under naive top-1 selection.

1.1 Related Literature

This paper contributes to several strands of the literature on post-selection inference, multiple comparison procedures, and machine learning for causal inference.

The methodological foundation is Generic Machine Learning (Chernozhukov et al., 2025a), which provides a framework for inference on heterogeneous treatment effects that is agnostic to the choice of ML learner. Chernozhukov et al. (2025a) acknowledge that the question of which learner to report remains open, suggesting either the union of confidence intervals or a Bonferroni correction when two learners are close in performance. Wager (2025) explores whether single-split and sequential procedures can reduce the computational burden of repeated splitting, while Imai and Li (2025) propose a cross-fitting approach based on randomization inference that requires only a single partition of the data. In their rejoinder, Chernozhukov et al. (2025b) demonstrate that the multi-split approach offers lower estimation risk and greater robustness to data mining relative to single-split alternatives. The present paper takes the multi-split framework as given and addresses the complementary problem of selecting among learners within it.

The learner selection problem in GML can be viewed as an instance of inference after selection, a topic that has received substantial attention in the statistics literature. Berk et al. (2013) develop valid post-selection inference for linear models, providing confidence intervals that account for all possible model selection procedures applied to the data. Fithian et al. (2017) characterize optimal inference after model selection by conditioning on the selection event, establishing a general framework based on sufficient statistics and the selection-adjusted likelihood. Andrews et al. (2024) study inference on the "winner" among competing estimates, deriving confidence intervals for the parameter associated with the largest observed value. Their setting is closely related to the GML learner selection problem: the researcher observes multiple estimates (one per learner), selects the best, and wishes to perform valid inference on the selected parameter. Zrnic and Fithian (2024) propose the zoom correction, a flexible method for inference on the winner that adapts to the gap between the best and remaining candidates, yielding tighter intervals when a clear winner emerges. I explored the zoom correction in the GML context but found that the power loss in low-signal settings was substantial, and the single-split requirement sacrificed the robustness benefits of median aggregation.

The model confidence set of Hansen et al. (2011) provides a different approach to selection. Rather than correcting inference for a selected model, the MCS constructs a set of models whose performances are statistically indistinguishable, in the same spirit as a confidence interval for a parameter. The MCS has been widely applied in forecasting and volatility model comparison (see, e.g., Hansen et al., 2011, and references therein), but has not previously been applied to learner selection in causal inference settings. The multiple testing procedures of Romano and Wolf (2007) offer a related framework for controlling familywise error rates across simultaneous hypotheses using resampling methods. While a Romano-Wolf stepdown procedure could in principle be applied to

the GML learner comparison problem, doing so requires estimating the joint covariance of median-aggregated test statistics across learners, which presents difficulties when influence functions are only valid conditional on their respective sample splits. The MCS sidesteps this issue by operating directly on split-level performance values.

2 Theory

The main focus of this paper is an application of the model confidence set (MCS) to learner results in Generic Machine Learning (GML) of Chernozhukov et al. (2025a). Hansen et al. (2011) does not make any assumptions about a model's relationship to the underlying DGP, rather, the model confidence set relies on a user-specified loss criterion of sample performance. The MCS procedure begins with a (large) set of candidate models, and removes candidates that are "not as good"–based on an equivalence test and and elimination rule–until the set of models is pruned to a set of models whose sample performances cannot be distinguished from one another for a given significance level. MCS has been applied in many ML settings involving time series data and forecasting (e.g., (Goulet Coulombe et al., 2022)), but has not been a popular approach for learner performance in causal inference settings. Although MCS was developed in the context of time series models, in this section I show that the MCS procedure–with a few tweaks–can be applied to the GML learner-selection problem.

2.1 Generic Machine Learning

GML offers a unique approach to estimating key features of the CATE, which happens in two stages which we might call *learning* and *post-processing*. Before learning, data $\mathcal{D} = (Y_i, D_i, Z_i)_{i=1}^N$, where Y_i is the outcome, D_i is a binary treatment indicator, and Z_i is a vector of covariates, are partitioned into an auxilliary sample $\mathcal{D}_A$ and a main sample $\mathcal{D}_M$. It may be more convenient for some readers to think of $\mathcal{D}_A$ and $\mathcal{D}_M$ as training and evaluation samples, respectively. Following Chernozhukov et al. (2025a), the partitions are assumed to be of roughly equal size ($N/2$).

Stage 1 (Learning): Let $Y(0)$ denote the outcome under the treated state and $Y(0)$ under non-treated. Using $\mathcal{D}_A$, estimate the baseline conditional average $b_0(Z) = \mathbb{E}[Y(0) \mid Z]$ and CATE $\mathbb{E}[Y(1) - Y(0) \mid Z]$ using ML methods (i.e., learners) $\ell = 1, ..., L$. Let the learned *proxy predictors* for the BCA and CATE (using $\mathcal{D}_A$) be denoted $B(Z)$ and $S(Z)$, respectively.

Stage 2 (Post processing): I consider two inferential targets in the post-processing stage for

the MCS procedure. First, the best linear predictor (BLP) of $s_0(Z)$ by the proxy $S(Z)$:

$$BLP[s_0(Z) \mid S(Z)] = \mathbb{E}[s_0(Z)] + \frac{\text{Cov}[s_0(Z), S(Z)]}{\text{Var}[S(Z)]}(S(Z) - \mathbb{E}[S(Z)]) \tag{1}$$

$$= \beta_1 + \beta_2 \, (S(Z) - \mathbb{E}[S(Z)]). \tag{2}$$

The BLP is identified by the weighted residual linear projection

$$Y = \alpha_0' X_1 + \beta_1(D - p(Z)) + \beta_2(D - p(Z))(S(Z) - \mathbb{E}[S(Z)]) + \epsilon, \quad \mathbb{E}[w(Z)\epsilon X] = 0, \tag{3}$$

where $p(Z) = \Pr(D = 1 \mid Z)$ is the propensity score, $w(Z) = \{p(Z)(1-p(Z))\}^{-1}$, and X_1 includes a constant, $B(Z)$, and other baseline controls. In a randomized experiment with constant propensity p, the weights simplify to a constant. The parameter β_2 is of primary interest: $\beta_2 = 1$ indicates that $S(Z)$ perfectly proxies the CATE, $\beta_2 = 0$ indicates that $S(Z)$ is uncorrelated with $s_0(Z)$ (or that $s_0(Z)$ is constant), and intermediate values reflect partial proxy quality. Note that if no heterogeneity is present in the true CATE, then β_2 is zero. Thus, a hypothesis test of $\beta_2 = 0$ is often utilized in practice as a test against zero heterogeneity. Using $\mathcal{D}_M$, one estimates $\hat{\beta}_1, \hat{\beta}_2$ and associated standard errors via OLS on (3) with $S(Z)$ replaced by $\hat{S}(Z)$ from the learning stage.

Second, sorted group average treatment effects (GATES). Partition observations in $\mathcal{D}_M$ into G groups based on the ranking of the proxy $\hat{S}(Z)$. In the simplest case with $G = 2$, let $G_k = \mathbb{1}\{\hat{S}(Z) \in \text{quantile group } k\}$ for $k = 1, 2$. The GATES are the group-specific average treatment effects $\gamma_k = \mathbb{E}[Y(1) - Y(0) \mid G_k = 1]$, estimated via the regression

$$Y = \alpha_0' X_1 + \sum_{k=1}^{G} \gamma_k (D - p(Z)) G_k + \nu, \quad \mathbb{E}[w(Z)\nu X] = 0. \tag{4}$$

The contrast $\gamma_G - \gamma_1$ tests whether the group predicted to benefit most from treatment actually has a larger treatment effect than the group predicted to benefit least, and serves as a direct test of whether $\hat{S}(Z)$ captures meaningful heterogeneity. As with the BLP, estimation proceeds by OLS on $\mathcal{D}_M$ with learned proxies from $\mathcal{D}_A$.

Chernozhukov et al. (2025a) emphasize that their methodology is generic with respect to the choice of machine learners used to form proxy predictors, and that $B(Z)$ and $S(Z)$ are generally noisy and inconsistent (without imposing untestable assumptions).[1] They navigate around the issue of poentailly-biased and noisy predictors via repeated sample splitting. Independently generated partitions $w = 1, \ldots, W$ yield split-specific estimates $\hat{\theta}_{\ell w}$ and standard errors $\hat{\text{se}}_{\ell w}$ for each learner ℓ and target parameter θ (e.g., β_2 or the contrast $\gamma_G - \gamma_1$). Within each split, inference conditional on the auxiliary sample is approximately Gaussian by Lemma D.1 of Chernozhukov et al. (2025a).

[1] See section 2.2-2.4, and the footnotes therein, of Chernozhukov et al. (2025a) for further explanation and a review of the issues faced in the ML + causal inference literature.

Across splits, point estimates and confidence bounds are median-aggregated, yielding valid unconditional inference that is robust to the randomness of any single partition. Specifically, within-split confidence intervals at nominal level ζ produce median-aggregated intervals with coverage at least $1 - 2\zeta$ (Theorems 4.2–4.3 of Chernozhukov et al. (2025a)).

2.2 The Learner Selection Problem

Each learner ℓ produces a different proxy $\hat{S}_\ell(Z)$, and consequently different estimates of the inferential targets $\hat{\theta}_\ell$. To compare learners, Chernozhukov et al. (2025a) introduce the goodness-of-fit functional

$$\Lambda(\ell) = |\beta_{2,\ell}|^2 \operatorname{Var}(S_\ell(Z)), \tag{5}$$

which can be interpreted as a measure of explained variation in the CATE attributable to proxy ℓ. Its sample analog, computed within each split, is $\hat{\Lambda}_w(\ell) = |\hat{\beta}_{2,\ell w}|^2 \widehat{\operatorname{Var}}[\hat{S}_{\ell w}(Z) \mid \mathcal{D}_M]$, and the median-aggregated version is $\tilde{\Lambda}(\ell) = \operatorname{med}_{w \in \mathcal{W}} \hat{\Lambda}_w(\ell)$.

A natural question arises: which learner's results should the researcher report? When a single learner clearly dominates the $\tilde{\Lambda}$ ranking, one may report that learner's inference with confidence. However, the apparent ranking may reflect finite-sample noise rather than genuine differences in population Λ values, particularly when many learners achieve similar performance. Chernozhukov et al. (2025a) acknowledge this issue and suggest two remedies for the case of two "near-winners": reporting the union of their confidence intervals or applying a Bonferroni correction. However, neither the determination of which learners qualify as near-winners nor the extension to more than two near-winners is formalized. The next section proposes a data-driven solution to both problems via the model confidence set.

Formalizing the near-winner set requires comparing learner performance in a way that accounts for the multiplicity of comparisons and the dependence structure induced by evaluating all learners on the same data. Two natural approaches present themselves, but each faces obstacles in the GML setting. First, one could test pairwise hypotheses of the form $H_0 : \Lambda(\ell) = \Lambda(\ell')$ using a studentized test statistic. However, while Lemma D.1 of Chernozhukov et al. (2025a) establishes the asymptotic normality of the split-sample estimators $\hat{\theta}_{\ell w}$ conditional on the auxiliary sample, no analogous distributional result exists for $\hat{\Lambda}_w(\ell)$. The goodness-of-fit functional involves a squared coefficient and a variance term whose joint distribution—and dependence across learners—is not characterized by the existing theory. Second, one could apply a Romano-Wolf stepdown procedure (Romano and Wolf, 2007) directly to the parameter estimates across learners. This requires estimating the joint covariance of the median-aggregated test statistics, which in turn involves pooling influence functions across splits that are each valid only conditional on their respective auxiliary samples. I instead take a different approach, treating the split-level performance values as a panel of loss observations and applying the model confidence set procedure of Hansen et al. (2011). The MCS

avoids both difficulties: it operates on the split-level $\hat{\Lambda}_w(\ell)$ values directly, and requires only that the loss differentials across splits satisfy a law of large numbers and central limit theorem. As shown below, this requirement is satisfied by the structure of GML's repeated sample splitting.

3 Model Confidence Set for GML

3.1 Setup and Definitions

Consider the set of learners $\mathcal{L} = \{1, \ldots, L\}$ and a collection of W independent random partitions $\mathcal{W} = \{1, \ldots, W\}$ of the data $\mathcal{D}$ into auxiliary and main samples $(\mathcal{D}_{A,w}, \mathcal{D}_{M,w})$ of equal size. For each split w and learner ℓ, the GML procedure yields a split-level goodness-of-fit value $\hat{\Lambda}_w(\ell)$ as defined in (5). The population counterpart is $\Lambda(\ell) = |\beta_{2,\ell}|^2 \operatorname{Var}(S_\ell(Z))$, and the objective is to determine which learners maximize $\Lambda(\cdot)$.

Define the set of superior learners as

$$\mathcal{M}^* = \{\ell \in \mathcal{L} : \Lambda(\ell) \geq \Lambda(j) \text{ for all } j \in \mathcal{L}\}. \tag{6}$$

When a unique best learner exists, $\mathcal{M}^*$ is a singleton; when multiple learners achieve the same population performance, $|\mathcal{M}^*| > 1$.

To apply the MCS framework, define the loss associated with learner ℓ in split w as the negative of the bias-corrected goodness-of-fit:

$$L_{\ell w} = -\hat{\Lambda}_w^{bc}(\ell), \qquad \hat{\Lambda}_w^{bc}(\ell) = \left(\hat{\beta}_{2,\ell w}^2 - \hat{\mathrm{se}}_{\ell w}^2\right) \hat{V}_{\ell w}, \tag{7}$$

where $\hat{V}_{\ell w} = \widehat{\operatorname{Var}}(\hat{S}_{\ell w}(Z) \mid \mathcal{D}_{M,w})$ is the sample variance of the proxy predictions in the evaluation fold. The negation ensures that learners with higher Λ have lower loss, so that the MCS procedure—which retains models with the smallest loss—selects learners with the best goodness of fit. The bias correction $\hat{\beta}_{2,\ell w}^2 - \hat{\mathrm{se}}_{\ell w}^2$ removes the leading-order upward bias in $\hat{\beta}_{2,\ell w}^2$ that arises because $\mathbb{E}[\hat{\beta}_2^2] = \beta_2^2 + \mathrm{se}^2$.

The relative performance of learner i versus learner j in split w is

$$d_{ij,w} = L_{i,w} - L_{j,w} = \hat{\Lambda}_w^{bc}(j) - \hat{\Lambda}_w^{bc}(i), \tag{8}$$

with population counterpart $\mu_{ij} = \mathbb{E}[d_{ij,w} \mid \mathcal{D}]$. Learner j is preferred to learner i when $\mu_{ij} < 0$, meaning learner i has higher expected loss. The null hypothesis of equal performance within a set $\mathcal{M} \subseteq \mathcal{L}$ is

$$H_{0,\mathcal{M}} : \mu_{ij} = 0 \quad \text{for all } i, j \in \mathcal{M}. \tag{9}$$

3.2 Equivalence Test

For a candidate set $\mathcal{M}$ with $|\mathcal{M}| = m$, define the relative sample loss of learner i against the average in $\mathcal{M}$:

$$\bar{d}_{i\cdot} = \frac{1}{W}\sum_{w=1}^{W} d_{i\cdot,w}, \qquad d_{i\cdot,w} = L_{i,w} - \frac{1}{m}\sum_{j\in\mathcal{M}} L_{j,w}, \tag{10}$$

with estimated variance $\widehat{\mathrm{Var}}(\bar{d}_{i\cdot}) = W^{-1}\cdot(W-1)^{-1}\sum_{w=1}^{W}(d_{i\cdot,w} - \bar{d}_{i\cdot})^2$. The t-statistic for learner i is

$$t_{i\cdot} = \frac{\bar{d}_{i\cdot}}{\sqrt{\widehat{\mathrm{Var}}(\bar{d}_{i\cdot})}}, \tag{11}$$

and the equivalence test statistic is

$$T_{\max,\mathcal{M}} = \max_{i\in\mathcal{M}} t_{i\cdot}. \tag{12}$$

A large value of $T_{\max,\mathcal{M}}$ indicates that at least one learner in $\mathcal{M}$ has significantly higher loss than the rest, providing evidence against $H_{0,\mathcal{M}}$.

3.3 Bootstrap Implementation

Critical values for $T_{\max,\mathcal{M}}$ are obtained via the nonparametric bootstrap. Since the split-level loss observations $\{L_{\ell w}\}_{w=1}^{W}$ are conditionally independent across w given $\mathcal{D}$, the standard i.i.d. bootstrap is appropriate. For each draw $b = 1, \ldots, B$:

1. Resample W split indices with replacement from $\{1, \ldots, W\}$.
2. Compute $d^*_{i\cdot,w}$ from the resampled loss matrix, then recenter by subtracting column means to impose $H_{0,\mathcal{M}}$.
3. Compute $t^*_{i\cdot}$ and $T^*_{\max,\mathcal{M}} = \max_{i\in\mathcal{M}} t^*_{i\cdot}$.

The p-value for the equivalence test on $\mathcal{M}$ is

$$\hat{p}_{H_{0,\mathcal{M}}} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left(T^{*(b)}_{\max,\mathcal{M}} \geq T_{\max,\mathcal{M}}\right). \tag{13}$$

3.4 The MCS Algorithm

The model confidence set $\hat{\mathcal{M}}_{\alpha_1}$ is constructed by the following sequential procedure at significance level α_1:

Step 0. Set $\mathcal{M} = \mathcal{L}$.
Step 1. Test $H_{0,\mathcal{M}}$ using $T_{\max,\mathcal{M}}$ at level α_1.

Step 2. If $H_{0,\mathcal{M}}$ is not rejected, set $\hat{\mathcal{M}}_{\alpha_1} = \mathcal{M}$ and stop. If $H_{0,\mathcal{M}}$ is rejected, eliminate the learner $e_{\mathcal{M}} = \arg\max_{i \in \mathcal{M}} t_{i\cdot}$ and return to Step 1 with the reduced set.

The elimination rule $e_{\mathcal{M}} = \arg\max_{i \in \mathcal{M}} t_{i\cdot}$ removes the learner contributing most to the rejection, and satisfies the coherency condition of Hansen et al. (2011) (Proposition 1) with the $T_{\max,\mathcal{M}}$ test.

Each learner ℓ receives an MCS p-value $\hat{p}_\ell = \max_{k \le k(\ell)} \hat{p}_{H_{0,\mathcal{M}_k}}$, where $k(\ell)$ is the elimination step at which ℓ is removed (or the final step if ℓ survives). By convention, the last surviving learner receives $\hat{p}_\ell = 1$. Learner ℓ belongs to $\hat{\mathcal{M}}_{\alpha_1}$ if and only if $\hat{p}_\ell > \alpha_1$.

3.5 Two-Stage Inference

With the model confidence set in hand, inference on the target parameter θ_ℓ (e.g., $\beta_{2,\ell}$ or a GATES contrast) proceeds over $\hat{\mathcal{M}}_{\alpha_1}$. For each learner $\ell \in \hat{\mathcal{M}}_{\alpha_1}$, construct within-split confidence intervals at nominal level $\zeta = \alpha_2/2$ and median-aggregate across splits:

$$\tilde{I}_{\theta_\ell}(\alpha_2) = \Big[\operatorname{med}_{w \in \mathcal{W}} \left(\hat{\theta}_{\ell w} - z_{1-\alpha_2/4}\, \hat{\mathrm{se}}_{\ell w} \right),\ \operatorname{med}_{w \in \mathcal{W}} \left(\hat{\theta}_{\ell w} + z_{1-\alpha_2/4}\, \hat{\mathrm{se}}_{\ell w} \right) \Big]. \tag{14}$$

By Theorems 4.2–4.3 of Chernozhukov et al. (2025a), each learner's median-aggregated interval covers θ_ℓ with probability at least $1 - \alpha_2 - o(1)$.

The union confidence interval over the model confidence set is

$$\tilde{I}^U(\alpha_2) = \left[\min_{\ell \in \hat{\mathcal{M}}_{\alpha_1}} \tilde{l}_{\theta_\ell},\ \max_{\ell \in \hat{\mathcal{M}}_{\alpha_1}} \tilde{u}_{\theta_\ell} \right], \tag{15}$$

where $\tilde{l}_{\theta_\ell}$ and $\tilde{u}_{\theta_\ell}$ denote the lower and upper bounds of $\tilde{I}_{\theta_\ell}(\alpha_2)$. The union interval has an oracle-best interpretation: if the best learner $\ell^* \in \mathcal{M}^*$ is contained in $\hat{\mathcal{M}}_{\alpha_1}$, then $\tilde{I}_{\theta_{\ell^*}}(\alpha_2) \subseteq \tilde{I}^U(\alpha_2)$, so the union interval covers θ_{ℓ^*} whenever the individual interval does. That is,

$$\Pr\Big(\theta_{\ell^*} \in \tilde{I}^U(\alpha_2) \Big) \ge \Pr\Big(\ell^* \in \hat{\mathcal{M}}_{\alpha_1} \text{ and } \theta_{\ell^*} \in \tilde{I}_{\theta_{\ell^*}}(\alpha_2) \Big). \tag{16}$$

Combining the MCS coverage guarantee with the median-aggregated coverage:

$$\Pr\Big(\theta_{\ell^*} \in \tilde{I}^U(\alpha_2) \Big) \ge 1 - \alpha_1 - \alpha_2 - o(1). \tag{17}$$

The first source of error (α_1) corresponds to the MCS failing to retain ℓ^*, and the second (α_2) to the individual confidence interval failing to cover θ_{ℓ^*}. Note that the union interval does not require a Bonferroni correction over $|\hat{\mathcal{M}}_{\alpha_1}|$ learners—each learner's interval is constructed at the unadjusted level α_2, and coverage of the oracle-best parameter follows from containment of a single interval within the union.[2]

[2]If simultaneous coverage of all θ_ℓ, $\ell \in \hat{\mathcal{M}}_{\alpha_1}$, is desired rather than coverage of the oracle best, one may replace

4 Validity of the MCS for GML

The MCS procedure of Section 3 inherits its theoretical guarantees from Hansen et al. (2011) provided that two conditions are met: (i) the loss differentials satisfy the regularity conditions required by the equivalence test and bootstrap, and (ii) the loss function correctly ranks learners in population, so that the MCS null hypothesis $H_{0,\mathcal{M}}$ holds when it should. The first condition is straightforward in the GML setting; the second requires care.

4.1 Regularity of the Loss Differentials

Assumption 2 of Hansen et al. (2011) requires the loss differentials $\{d_{ij,w}\}_{w=1}^{W}$ to be strictly stationary, α-mixing, and to possess finite $(2+\gamma)$-moments. In the GML setting, the partitions $(\mathcal{D}_{A,w}, \mathcal{D}_{M,w})$ are drawn independently across w, conditional on $\mathcal{D}$. Therefore:

- **Independence.** The loss observations $\{L_{\ell w}\}_{w=1}^{W}$ are conditionally i.i.d. given $\mathcal{D}$. Strict stationarity and α-mixing of any order are trivially satisfied, as i.i.d. sequences are a special case.
- **Finite moments.** Each $\hat{\Lambda}_w^{bc}(\ell)$ is a product of OLS quantities computed on a sample of size $N/2$. Under standard regularity conditions on the design matrix (bounded eigenvalues, finite fourth moments of the data), the estimators $\hat{\beta}_{2,\ell w}$, $\hat{\mathrm{se}}_{\ell w}$, and $\hat{V}_{\ell w}$ have finite moments of all orders, and so does $d_{ij,w}$.
- **Positive variance.** The requirement $\mathrm{Var}(d_{ij,w}) > 0$ holds as long as distinct learners do not produce identical predictions in every split, which is guaranteed whenever the learners employ different algorithms or tuning parameters.

With Assumption 2 satisfied, Theorem 4 of Hansen et al. (2011) establishes that the equivalence test based on $T_{\max,\mathcal{M}}$ has the correct asymptotic size under $H_{0,\mathcal{M}}$ and diverges under the alternative, and that the elimination rule $e_{\mathcal{M}} = \arg\max_{i\in\mathcal{M}} t_i$. removes an inferior learner with probability approaching one. Proposition 1 of Hansen et al. (2011) guarantees coherency between this test-elimination pair. Together, these results verify Assumption 1 of Hansen et al. (2011), and Theorem 1 then delivers the coverage guarantee $\Pr(\mathcal{M}^* \subseteq \hat{\mathcal{M}}_{\alpha_1}) \geq 1-\alpha_1$.

The bootstrap is consistent by standard arguments for i.i.d. data. Recentering the resampled loss differentials by their column means imposes $H_{0,\mathcal{M}}$ in the bootstrap world, following the unrestricted bootstrap approach of Hansen et al. (2011, Section 3.1) and White (2000).

$\tilde{I}_{\theta_\ell}(\alpha_2)$ with Bonferroni-corrected intervals at level $\alpha_2/|\hat{\mathcal{M}}_{\alpha_1}|$.

4.2 Validity of the Loss Function

The regularity conditions above ensure that the MCS procedure works correctly *given* a valid loss function. The remaining question is whether the bias-corrected Lambda defined in (7) satisfies the key requirement: under the null that learners i and j have equal population performance, $\Lambda(i) = \Lambda(j)$, the expected loss differential must vanish,

$$\mathbb{E}[d_{ij,w} \mid \mathcal{D}] = 0 \quad \text{when} \quad \Lambda(i) = \Lambda(j). \tag{18}$$

This is not automatic. The bias-corrected goodness-of-fit is $\hat{\Lambda}_w^{bc}(\ell) = (\hat{\beta}_{2,\ell w}^2 - \hat{\text{se}}_{\ell w}^2)\hat{V}_{\ell w}$. Taking conditional expectations and decomposing:

$$\mathbb{E}[\hat{\Lambda}_w^{bc}(\ell) \mid \mathcal{D}] = \underbrace{\mathbb{E}[\hat{\beta}_{2,\ell w}^2 - \hat{\text{se}}_{\ell w}^2 \mid \mathcal{D}]}_{\approx \beta_{2,\ell}^2} \cdot \underbrace{\mathbb{E}[\hat{V}_{\ell w} \mid \mathcal{D}]}_{\approx \text{Var}(S_\ell(Z))} + \underbrace{\text{Cov}\Big(\hat{\beta}_{2,\ell w}^2 - \hat{\text{se}}_{\ell w}^2,\ \hat{V}_{\ell w} \mid \mathcal{D}\Big)}_{\text{cross-term}}. \tag{19}$$

The first two factors deliver $\Lambda(\ell)$ as desired. The concern is the cross-term: $\hat{\beta}_{2,\ell w}$ and $\hat{V}_{\ell w}$ are both estimated on $\mathcal{D}_{M,w}$ and are linked through the proxy $\hat{S}_\ell(Z)$, which enters both the BLP regressor and the variance calculation. If this cross-term differs systematically across learners, the centering condition (18) fails and the MCS may eliminate learners based on an artifact rather than genuine performance differences.

The following argument shows that the cross-term is asymptotically negligible. Conditional on $\mathcal{D}_{A,w}$, the proxy $\hat{S}_\ell(Z)$ is a fixed, nonrandom function. The sample variance $\hat{V}_{\ell w}$ is then a standard sample statistic evaluated at the $N/2$ observations in $\mathcal{D}_{M,w}$, satisfying

$$\hat{V}_{\ell w} = \text{Var}(S_\ell(Z)) + O_p(N^{-1/2}). \tag{20}$$

The covariance of any $O_p(1)$ quantity with an $O_p(N^{-1/2})$ quantity is $O(N^{-1/2})$. Therefore the cross-term in (19) vanishes at rate $O(N^{-1/2})$, uniformly across learners. The resulting distortion to the t-statistic in (11) is of order $O(\sqrt{W/N})$, which is negligible provided $W = o(N)$. Under the configurations considered in this paper ($N = 1000$, $W \leq 101$), this condition is comfortably satisfied, and the simulation results in Section 5 confirm that the MCS achieves correct size control under the null.

Additional work on this section is needed

5 Simulation Results

The data generating process follows Wager (2025) and the rejoinder of Chernozhukov et al. (2025b). Covariates $Z_i \in \mathbb{R}^{10}$ are drawn i.i.d. from Uniform$(-1, 1)^{10}$, treatment is assigned as $D_i \sim$ Bernoulli(0.5),

and outcomes are generated as

$$Y_i = D_i \cdot \tau(Z_i) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0,1), \tag{21}$$

with $\tau(z) = \max(0, z_1)$ under the heterogeneous design and $\tau(z) = 0$ under the homogeneous design. The sample size is $n = 1000$. This DGP provides a useful benchmark: the CATE is a simple, piecewise-linear function of a single covariate, producing moderate treatment effect heterogeneity that is detectable but not overwhelming. Only one of the ten covariates is relevant, creating a setting in which learner performance is expected to vary meaningfully across ML methods.

I run 2000 Monte Carlo replications for both homogeneous and heterogeneous versions of (21) and record rejection rates from a test $H_0 : \beta_2 = 0$, where β_2 is the parameter in (3). Recall that $s_0(Z)$ is constant under the homogeneous DGP, and thus $\text{Cov}[s_0(Z), S(Z)] \approx 0$ which implies failure to reject $\beta_2 = 0$. Under the heterogeneous DGP, the theoretical value for β_2 is unknown. However, because heterogeneity is present, a good proxy will uncover the systematic variation in $s_0(Z)$ and result in a nonzero estimate of β_2.[3]

A total of seven learners were used: partial least squares, random forest, lasso, trees, generalized linear model, k-nearest neighbors, and causal forest. All are implemented in `R` via the `caret` package (Kuhn and Max, 2008) except causal forest, for which I use the `grf` package (Tibshirani et al., 2024; Athey et al., 2019).[4] For all but causal forest, the CATE predictions are generated via the "T-learner" approach (see Künzel et al. (2019) for more details on meta learners). This involves simply fitting ℓ to $\mathcal{D}_A \mid D_i = 0$ to obtain $\mu_{(0)\ell}(Z)$ and to $\mathcal{D}_A \mid D_i = 1$ to obtain $\mu_{(1)\ell}(Z)$.[5] CATE predictions are obtained by generating predictions of both $\mu_{(0)\ell}(Z)$ and $\mu_{(1)\ell}(Z)$ in the $\mathcal{D}_M$ sample and taking the difference, $S_\ell(Z) = \hat{\mu}_{(1)\ell}(Z) - \hat{\mu}_{(0)\ell}(Z)$.[6] Causal forests do not require the T-learner as the method is purpose-built for learning the CATE. As mentioned in the introduction, it is rarely the case that a researcher has an *a priori* about which learning algorithms are best suited to uncovering patterns in the true CATE.[7] Beginning with a relatively large set of candidate learners casts a wide net and offers a prudent attempt at uncovering these patterns.

To illustrate the challenge of learner selection, Table 1 reports the mean rank and standard deviation of each learner's bias-corrected $\tilde{\Lambda}^{bc}$ across Monte Carlo replications. Under the homogeneous DGP, all mean ranks cluster near the midpoint of 4 with standard deviations around 2, indicating that the ranking is effectively random. A researcher selecting the "best" learner in this setting is choosing based on noise. Under the heterogeneous DGP, a hierarchy emerges: lasso and

[3]The sign of β_2 is positive given a good proxy since $\text{Cov}[s_0(Z), S(Z)] < 0$ would be a case where $S(Z)$ predicts heterogeneity in the wrong direction.

[4]For caret learners: the regression versions of `pls, rf, lasso, rpart, glm, knn`. `grf::causal_forest()` is used to implement the causal forest of Athey et al. (2019).

[5]Note that $\mu_{(0)\ell}(Z)$ is the BCA, $B_\ell(Z)$.

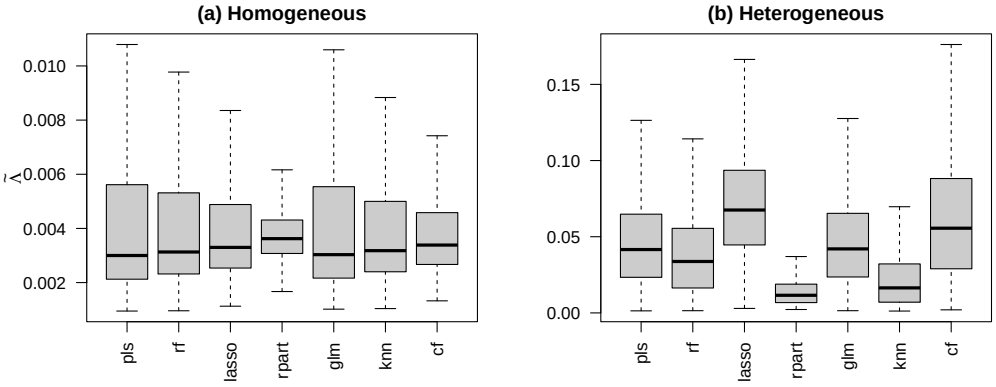
[6]The T-learner is a common method employed in practice and the one utilized in the popular `R` package, `GenericML` (Welz et al., 2022).

[7]One notable exception is Deryugina et al. (2019).

Table (1) Mean rank and standard deviation of $\tilde{\Lambda}^{bc}$ rank by learner, DGP 3. Rank 1 corresponds to the highest $\tilde{\Lambda}^{bc}$.

	Homogeneous		Heterogeneous	
	Mean	SD	Mean	SD
pls	4.18	2.02	3.71	1.08
rf	4.10	2.21	4.41	1.40
lasso	3.99	1.83	1.41	0.72
rpart	3.70	1.94	6.49	0.85
glm	4.21	1.99	3.56	1.12
knn	4.06	2.22	5.96	1.08
cf	3.76	1.68	2.46	1.29

Figure (1) Distribution of bias-corrected $\tilde{\Lambda}^{bc}$ by learner across Monte Carlo replications

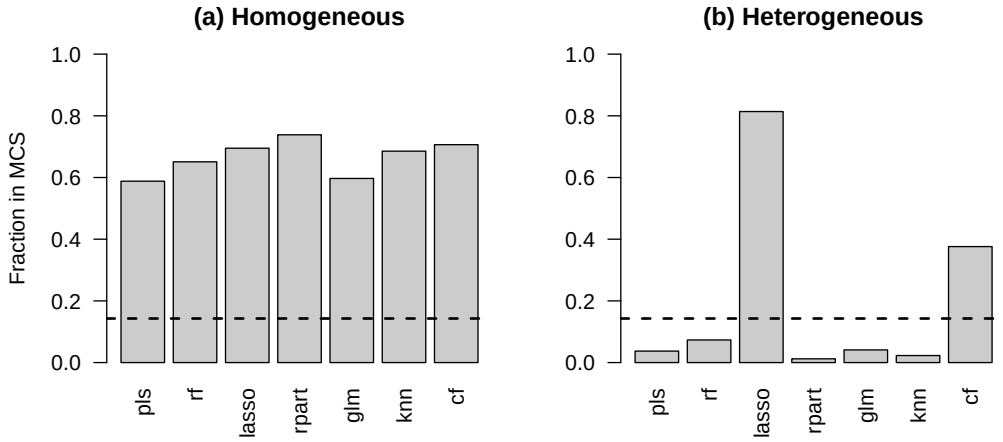


causal forest rank highest on average, but with sufficient variability that the identity of the second- or third-best learner shifts across replications.

Figure 1 displays the distribution of $\tilde{\Lambda}^{bc}$ by learner. Under the homogeneous design, all learners produce indistinguishable performance values. Under the heterogeneous design, lasso and causal forest separate from the remaining learners, but the overlap in distributions underscores the difficulty of drawing a sharp boundary between near-winners and clearly inferior candidates.

Table 3 displays the simulation results. Size and power correspond to rejection rates under the homogeneous and heterogeneous processes, respectively. The first row contains the results from selecting the top-performing learner based on the standard median-aggregated $\hat{\Lambda}(\ell)$. Size is held under the nominal level, and power is nearly 80 percent. If instead a researcher chose to select the top two learners as near-winners and make an *ad hoc* adjustment to the confidence interval bounds, the power drops nearly 16 percent with a Bonferroni correction, and nearly 26 percent

Figure (2) Distribution of bias-corrected $\tilde{\Lambda}^{bc}$ by learner across Monte Carlo replications



when taking the union of the bounds. The MCS procedure, by contrast, achieves a rejection rate of 76.4 percent under the heterogeneous DGP–a loss of only 3.2 percentage points relative to naive top-1 selection. This is notable given that the MCS retains an average of 1.38 learners under the alternative, which is comparable to the fixed top-2 selection. The key difference is adaptivity: the MCS selects $|\hat{\mathcal{M}}|= 1$ in 73 percent of replications, applying no correction at all, and expands to two or more learners only when the data cannot distinguish the best from its competitors. By contrast, the fixed top-2 rule always corrects for two learners regardless of whether the second-best learner is genuinely competitive. Under the homogeneous DGP, the MCS expands to a median of 5 learners with roughly uniform membership across all seven candidates, correctly reflecting that no learner outperforms any other when the CATE is constant.

Figure 2 displays the distribution of $\tilde{\Lambda}^{bc}$ by learner. Under the homogeneous design, all learners produce indistinguishable performance values. Under the heterogeneous design, lasso and causal forest separate from the remaining learners, but the overlap in distributions underscores the difficulty of drawing a sharp boundary between near-winners and clearly inferior candidates.

Table 2 reports the range of median-aggregated $\hat{\beta}_2$ estimates across learners retained by the MCS. Under the homogeneous DGP, the MCS retains multiple learners in 87.5 percent of replications, and the average range of point estimates across the surviving set is 0.54. Since the true β_2 is zero, this variation is entirely driven by estimation noise, yet all of these estimates are equally defensible given the MCS. A researcher reporting only the naive top-1 estimate would present a single number (on average, 0.024 above the truth) without any indication that comparably performing learners disagree by over half a unit. Under the heterogeneous DGP, the MCS retains multiple

Table (2) Sensitivity of $\hat{\beta}_2$ to learner choice within the MCS

	Homogeneous	Heterogeneous
Fraction $\lvert\hat{\mathcal{M}}\rvert > 1$	0.875	0.266
Mean range of $\hat{\beta}_2$ across $\hat{\mathcal{M}}$	0.470	0.166
Mean range when $\lvert\hat{\mathcal{M}}\rvert > 1$	0.537	0.622
Fraction with range > 0.1	0.823	0.262
Fraction with range > 0.5	0.450	0.165

Notes: Range is computed as $\max_{\ell \in \hat{\mathcal{M}}} \tilde{\theta}_\ell - \min_{\ell \in \hat{\mathcal{M}}} \tilde{\theta}_\ell$ where $\tilde{\theta}_\ell$ is the median-aggregated $\hat{\beta}_2$ for learner ℓ. When $\lvert\hat{\mathcal{M}}\rvert = 1$, range is zero.

learners less frequently (27 percent of replications), but when it does, the average range is 0.62. This variation reflects genuine differences in how learners approximate the CATE and has direct implications for applied work: two learners that the MCS deems statistically indistinguishable may produce meaningfully different estimates of the degree of treatment effect heterogeneity.

One might ask whether a simpler, threshold-based rule could achieve similar results–for example, defining the near-winner set as all learners within some fraction c of the maximum $\tilde{\Lambda}$. Such a rule requires the researcher to choose c, and the choice has direct consequences: too small and the set collapses to a singleton regardless of whether near-winners exist; too large and the set includes clearly inferior learners, unnecessarily widening confidence intervals. Moreover, the appropriate value of c depends on the dispersion of $\hat{\Lambda}$ across splits, which varies with the DGP, sample size, and learner set. The MCS replaces this subjective choice with a formal test calibrated at level α_1. When the data strongly separate learners, the equivalence test rejects repeatedly and the set shrinks; when learners are indistinguishable, the test accepts early and the set remains large. This behavior is governed entirely by the sampling variability of the split-level performance values, not by a tuning parameter.

6 Application to Loans in Egypt

This section applies the MCS to Bryan et al. (2024)–an analysis of a business-loan RCT in Egypt where the authors utilize GML to investigate heterogeneity in treatment effects.

Forthcoming

7 Conclusion

This paper proposes applying the model confidence set of Hansen et al. (2011) to the learner selection problem in Generic Machine Learning. The procedure provides a data-driven, formally justified

Table (3) Size and Power for BLP, $n = 1000$, $W = 101$, $\alpha_1 = \alpha_2 = 0.05$

Method	Hom (size)	Het (power)	Mean $\lvert\hat{\mathcal{M}}\rvert$	
			Hom	Het
Naive top-1, unadjusted	0.005	0.796	1	1
Naive top-2, BF(2)	0.002	0.683	2	2
Naive top-2, union	0.002	0.538	2	2
MCS + BF, best-in-$\hat{\mathcal{M}}$	0.004	0.764	4.67	1.38
MCS + union	0.004	0.755	4.67	1.38

Notes: 2000 Monte Carlo replications. Homogeneous DGP: $Y = \varepsilon$. Heterogeneous DGP: $Y = D \cdot \max(0, Z_1) + \varepsilon$. Learner set: pls, rf, lasso, rpart, glm, knn, cf. Bias-corrected Λ used for MCS and naive selection.

determination of which learners are statistically indistinguishable in their ability to proxy the CATE, replacing the ad hoc selection that currently characterizes applied practice. The resulting near-winner set feeds naturally into the multiplicity corrections already recommended by Chernozhukov et al. (2025a), with the key advantage that the size of the correction adapts to the informativeness of the data.

Three findings emerge from the Monte Carlo analysis. First, the two-stage procedure controls size at the nominal level across all specifications considered. Second, the power cost relative to naive selection is modest, owing to the MCS's ability to shrink to a singleton when a clear winner exists. Third, the MCS reveals meaningful variation in point estimates across equally credible learners, providing applied researchers with a diagnostic that is unavailable under standard reporting practices.

Several directions for future work remain. The current analysis focuses on the BLP parameter β_2; extending the MCS framework to GATES and CLAN parameters would broaden its applicability to policy-relevant questions about treatment targeting. The procedure is also compatible with the cross-fitting approach of Imai and Li (2025) and the sequential framework of Wager (2025), though formal integration with these alternatives requires additional theoretical development. Finally, the MCS approach extends naturally beyond GML to any setting where multiple machine learning methods produce competing causal estimates, including double/debiased machine learning and policy learning.

A Additional Results

Table (4) Mean $\tilde{\Lambda}$ and bias-corrected $\tilde{\Lambda}^{bc}$ by learner

	Homogeneous			Heterogeneous		
	$\tilde{\Lambda}$	$\tilde{\Lambda}^{bc}$	Diff.	$\tilde{\Lambda}$	$\tilde{\Lambda}^{bc}$	Diff.
pls	0.00502	-0.00308	0.00811	0.04751	0.03898	0.00853
rf	0.00474	-0.00338	0.00812	0.03897	0.03041	0.00856
lasso	0.00449	-0.00362	0.00811	0.07174	0.06320	0.00853
rpart	0.00378	-0.00434	0.00812	0.01416	0.00559	0.00856
glm	0.00499	-0.00311	0.00810	0.04795	0.03941	0.00854
knn	0.00469	-0.00343	0.00812	0.02217	0.01361	0.00856
cf	0.00424	-0.00381	0.00805	0.06255	0.05401	0.00854
Spearman ρ		0.90650			0.98940	

References

Andrews, Isaiah, Toru Kitagawa, and Adam McCloskey, "Inference on winners," *The Quarterly Journal of Economics*, 2024, *139* (1), 305–358.

Athey, Susan, Julie Tibshirani, and Stefan Wager, "Generalized random forests," *The Annals of Statistics*, 2019, *47* (2), 1148 – 1178.

Berk, Richard, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao, "Valid post-selection inference," *The Annals of Statistics*, 2013, pp. 802–837.

Bryan, Gharad, Dean Karlan, and Adam Osman, "Big loans to small businesses: Predicting winners and losers in an entrepreneurial lending experiment," *American Economic Review*, 2024, *114* (9), 2825–2860.

Chernozhukov, Victor, Mert Demirer, Esther Duflo, and Iván Fernández-Val, "Fisher–Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India," *Econometrica*, 2025, *93* (4), 1121–1164.

_ , _ , _ , and _ , "Reply to: Comments on "Fisher–Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India"," *Econometrica*, 2025, *93* (4), 1177–1181.

Coulombe, Philippe Goulet, Maxime Leroux, Dalibor Stevanovic, and Stéphane Surprenant, "How is machine learning useful for macroeconomic forecasting?," *Journal of Applied Econometrics*, 2022, *37* (5), 920–964.

Deryugina, Tatyana, Garth Heutel, Nolan H Miller, David Molitor, and Julian Reif, "The mortality and medical costs of air pollution: Evidence from changes in wind direction," *American Economic Review*, 2019, *109* (12), 4178–4219.

Fithian, William, Dennis Sun, and Jonathan Taylor, "Optimal Inference After Model Selection," 2017.

Hansen, Peter R, Asger Lunde, and James M Nason, "The model confidence set," *Econometrica*, 2011, *79* (2), 453–497.

Imai, Kosuke and Michael Lingzhi Li, "A Comment on: "Fisher–Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India" by Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val," *Econometrica*, 2025, *93* (4), 1165–1170.

Kuhn and Max, "Building Predictive Models in R Using the caret Package," *Journal of Statistical Software*, 2008, *28* (5), 1–26.

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

Romano, Joseph P. and Michael Wolf, "Control of generalized error rates in multiple testing," *The Annals of Statistics*, 2007, *35* (4), 1378 – 1408.

Tibshirani, Julie, Susan Athey, Erik Sverdrup, and Stefan Wager, *grf: Generalized Random Forests* 2024. R package version 2.4.0.

Wager, Stefan, "A Comment on: "Fisher–Schultz Lecture: Generic Machine Learning Inference on Heterogeneous Treatment Effects in Randomized Experiments, With an Application to Immunization in India" by Victor Chernozhukov, Mert Demirer, Esther Duflo, and Iván Fernández-Val," *Econometrica*, 2025, *93* (4), 1171–1176.

Welz, Max, Andreas Alfons, Mert Demirer, and Victor Chernozhukov, *GenericML: Generic Machine Learning Inference* 2022. R package version 0.2.2.

White, Halbert, "A reality check for data snooping," *Econometrica*, 2000, *68* (5), 1097–1126.

Zrnic, Tijana and William Fithian, "A Flexible Defense Against the Winner's Curse," 2024.