

Double/Debiased Machine Learning (DDML) for Conditional Average Treatment Effects (CATE) Estimation

Luiza Antonie, University of Guelph
Mohammad Havaledar, University of Guelph
Miana Plesca, University of Guelph

Carpathian Stata Conference
July 28, 2026 Cluj-Napoca

What is CATE? Conditional Average Treatment Effect

Athey and Imbens (2016), Wager and Athey (2018)

- Allows for **heterogeneity** in the impact of treatment

$$\tau_0(z) = E[Y(1) - Y(0) \mid Z = z]$$

- Y is outcome and Z are the variables used to describe heterogeneity in treatment effects.

For example, Y is income and Z can include age, education, race, earnings history, occupation, employer, region, etc.

- **Causal** in the sense that focus shifts away from predicting outcome to comparing potential outcomes under treatment and no treatment.

Needs CIA to hold for identification: $(Y(1), Y(0)) \perp D \mid X$ and $0 < P(D = 1 \mid X) < 1$.

- A good ATE is not enough if the research question is about who gains, who loses, and where effects are concentrated.

What is DDML? Double/Debiased Machine Learning

Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, and Robins (2018)

Chernozhukov, Escanciano, Ichimura, Newey, and Robins (2022)

- ML can introduce regularization and overfitting bias.
- **Orthogonalization** is used to reduce regularization bias from the ML first stages.
- **Cross-fitting** estimates nuisance models on one part of the data and evaluates the causal score on another part.

Ahrens, Hansen, Schaffer, Wiemann (2024) “ddml”

What is the effect of choice of Machine Learner and Hyper Parameter Optimization (HPO) when the target is CATE?

- A machine learner can predict the outcome well but still give a poor treatment-effect estimate.
- A learner can recover the average effect while missing treatment-effect heterogeneity.
- Tuning choices that look reasonable for prediction can create unstable participation probabilities or unstable estimated CATEs.

Partially Linear and Interactive CATE Regression Models

PLRM / partialing-out “cate po”

$$Y = \ell(X) + D(X)\tau(Z) + U, \quad D = m(X) + V.$$

- $m(X)$ is the participation model.
- $\ell(X)$ is the PLRM outcome nuisance.
- $CATE = \tau(Z)$ varies with Z .
- orthogonalization is through residuals on residuals regression.

IRM / AIPW “cate aipw”

$$Y = g(X, D) + U, \quad D = m(X) + V.$$

$$g_1(X) = E[Y \mid D = 1, X], \quad g_0(X) = E[Y \mid D = 0, X], \quad m(X) = P(D = 1 \mid X).$$

- $g(X)$ is the IRM/AIPW outcome nuisance.
- $CATE = g_1(Z) - g_0(Z)$ where Z is a subset of X .
- orthogonalization is through AIPW.

Experiment map: what is being compared?

Experiment block	Purpose	Main comparison
Synthetic DGPs	Known truth. Lets us directly evaluate CATE accuracy.	Linear vs nonlinear outcome and participation designs.
LaLonde/NSW-PSID	Experimental benchmark plus observational stress test.	RCT, PSID controls , all controls, placebo controls.

Do conclusions change when we move from known-truth simulations to a real observational benchmark?

Estimator	Purpose	Main comparison
Native Stata cate	Main DDML-style implementation. Built-in orthogonalization and cross-fitting.	cate po vs cate aipw Lasso vs RF-HPO.
Manual orthogonal learners	Diagnostics for cross-fitting.	PLR vs IRM cross-fitted vs not cross-fitted.
Non-DDML baselines	Non-orthogonal CATE reference points.	S-learner, T-learner.

When do orthogonalization and cross-validation matter in practice?

Five common CATE meta-learner families

Learner	Basic idea	Role in this deck
S-learner	Estimate one outcome model using treatment plus covariates, then compare predicted outcomes with treatment switched on/off.	non-DDML baseline
T-learner	Estimate separate outcome models for treated and untreated groups, then difference the predictions.	non-DDML baseline
X-learner	Impute treatment effects within treated and untreated groups, then combine them using a weighting rule.	taxonomy/reference, not central in code
R-learner	Residualize outcome and treatment, then learn treatment-effect heterogeneity from the residualized relationship.	close to PLR/manual orthogonal logic
DR-learner	Build a doubly robust pseudo-outcome and regress it on heterogeneity variables.	close to IRM/AIPW logic

The native `cate po` and `cate aipw` routines are the main DDML-style estimators here. The S/T/IPW learners are included as simpler benchmarks.

Experiment map: Limitations

- DGP has only $n = 4,000$. LaLonde NSW is also small n small p .
- No evaluation for p/n .
- Not sufficiently investigated: high-dimensional categorical X (e.g. occupation fixed effects).
- Only learners evaluated are: OLS, Lasso, RF. Other learners (gradient boosting, neural nets, etc.) can be more effective with large data.
- Combinations of different learners for outcome, participation.
- Choice of Z as subset of X .
- Loss functions in evaluating results: recent work on “grounded truth” DGP.

Synthetic DGPs: outcome, treatment-effect and participation equations

Four scenarios LL, LN, NL, NN: outcome(treatment) and participation

Common setup: $x_1, \dots, x_8$ are standard normal variables, $z_1 = 1[x_5 > 0]$, and $z_2 = 1[x_6 > 0]$. Each simulated dataset has $N = 4,000$.

DGP	Baseline outcome $\mu_0(X)$	Treatment effect $\tau(X)$
LL	$1 + .70x_1 - .60x_2 + .35x_3 + .25x_4 + .40x_5 - .30x_6 + .35z_1 - .25z_2$	$1 + .70x_1 - .40x_2 + .30x_3 - .20x_4$
LN	same linear $\mu_0(X)$ as LL	same linear $\tau(X)$ as LL
NL	$1 + \sin(x_1) + .60(x_2^2 - 1) + .50x_3x_4 + .30x_5 - .25x_6 + .70z_1 - .50z_2$	$1 + .80 \sin(x_1) - .45(x_2^2 - 1) + .65x_3x_4 + .50z_1 - .35z_2$
NN	same nonlinear $\mu_0(X)$ as NL	same nonlinear $\tau(X)$ as NL

Treatment is drawn from a logistic participation model. The true propensity is clipped to avoid a completely deterministic treatment rule.

Participation design	Index used inside the logit participation model	DGPs
Linear participation	$-.15 + .65x_1 - .55x_2 + .35x_3 + .20x_4 + .30z_1 - .20z_2$	LL, NL
Nonlinear participation	$-.15 + .85 \sin(x_1) - .70(x_2^2 - 1) + .55x_3x_4 + .35x_5 - .25x_6 + .45z_1 - .30z_2$	LN, NN

LL/LN are friendly to linear CATE models. NL/NN require learners to recover nonlinear heterogeneity.

Actual Stata commands behind the native CATE estimates

The generic native specification is:

```
cate po    ($Y $CATEVARS) ($D), controls($CONTROLS)  ///
      omethod(...) tmethod(...) cmethod(...)          ///
      xfolds(5) rseed(90210) nolog
```

```
cate aipw ($Y $CATEVARS) ($D), controls($CONTROLS)  ///
      omethod(...) tmethod(...) cmethod(...)          ///
      xfolds(5) rseed(90210) nolog
```

Examples of learner choices:

```
omethod(lasso, selection(plugin))
tmethod(lasso, selection(plugin))
cmethod(rforest, ntrees(600) splitminobs(10)
        splitmeanvars(4) samprate(.50))
```

- `omethod`: outcome nuisance learner.
- `tmethod`: treatment/participation learner.
- `cmethod`: CATE learner.
- Native cate always uses five-fold cross-fitting here.

Machine learners and parametrization

Learner family	Where used	Parametrization
Parametric	Native and manual baseline	OLS regression for outcome; logit for participation; OLS regression for CATE.
Lasso plug-in	Native and manual nuisance	<code>selection(plugin)</code> . Penalty chosen by theory-driven plug-in rule.
Lasso CV	Native and manual nuisance	<code>selection(cv, folds(5))</code> . Penalty chosen by five-fold CV.
Random forest HPO	Native CATE and nuisance	Five RF candidates over trees, split size, split variables, sampling rate, and honesty.
Non-DDML baselines	S/T/IPW-style comparisons	OLS, lasso, logit, and lasso-logit variants without orthogonal DDML structure.

Learner selection is part of the experiment: the same DGP can look different under parametric, lasso, and RF-HPO choices.

Specification spec and cross-fit xfit models

Label	Meaning	Important interpretation
<code>family = native_cate_hpo</code>	Stata native cate.	These are the blue rows/points. A row such as <code>lasso_cv_rfHPO_rf3_mid</code> is still native cate.
<code>family = manual_orth</code>	Hand-coded orthogonal PLR/IRM diagnostics.	These are orange. They are not Stata native cate.
<code>family = nonddml</code>	Non-orthogonal S/T/IPW baselines.	These are gray and are included as benchmarks.
<code>xfit = xfold5</code>	Five-fold cross-fitting.	This is DDML sample-splitting, not lasso HPO. Native cate uses <code>xfolds(5)</code> .
<code>xfit = noxfit</code>	No cross-fitting.	Used only for non-DDML and manual diagnostics.
<code>lasso_cv</code>	Lasso penalty selected by inner cross-validation.	This is lasso HPO.
<code>rfHPO_rf*</code>	Random forest candidate selected from the RF grid.	This is RF HPO; the final estimate is still native cate if <code>family = native_cate_hpo</code> .

Reading rule: spec describes learner/HPO choices. xfit describes cross-fitting. They are related but not the same.

Lasso tuning and Random forest HPO grid

Lasso tuning

```
lasso linear y x1 x2 ..., selection(plugin)
```

```
lasso linear y x1 x2 ..., selection(cv, folds(5))
```

- Cross-validated lasso chooses the penalty that predicts well in held-out folds.
- Plug-in lasso chooses the penalty using a theory-driven rule intended to control noise-variable selection.
- In practice, plug-in lasso is often more conservative than CV lasso. That can matter for DDML: nuisance prediction is not the same as causal accuracy!

RF tuning

- External to cate: a tuning subsample is used to select a candidate RF specification, then the selected RF is run on the full sample.

Candidate	trees	split min	split vars	samp rate	interpretation
RF1-cons	400	20	2	.50	conservative forest
RF2-small	600	10	4	.50	moderate forest
RF3-mid	800	6	5	.60	larger middle option
RF4-deep	1000	3	6	.70	deeper/aggressive option
RF5-aggr	1200	3	6	.75	aggressive, no-honesty option

Some RF candidates may fail in Stata because cate imposes restrictions on combinations such as `samprate()`

HPO design

- A 30 percent tuning subsample is created for RF HPO.
- In simulations, HPO can use the true CATE because the DGP is known.
- In LaLonde, HPO can use distance from the RCT benchmark only as a retrospective research exercise.
- In an ordinary observational application, this benchmark-guided tuning would not be available.

Data block	HPO criterion used in this run	Interpretation
Synthetic simulations	PEHE	Selects RF specification closest to true heterogeneity.
RCT LaLonde	distance from experimental benchmark	Ex-post benchmark exercise.
Observational LaLonde	distance from RCT benchmark	Method-evaluation exercise, not ordinary applied tuning.
Placebo controls	distance from zero	Selection diagnostic.

Evaluation metrics and where they are used

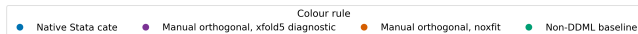
Measure	What it tells us	Simulation	RCT	Observational	Placebo
atehat	Estimated average effect from the CATE estimator.	✓	✓	✓	✓
trueate	True average effect from the DGP.	✓	–	–	–
bias_true	Difference between estimated and true ATE.	✓	–	–	–
pehe	Precision in Estimation of Heterogeneous Effect: CATE/IATE RMSE against true individual effects.	✓	–	–	–
corr	Correlation between estimated and true CATE.	✓	–	–	–
benchate	Experimental RCT benchmark.	–	✓	✓	✓
bias_bench	Difference between estimate and benchmark.	–	✓	✓	✓
tau_sd	Spread of estimated individual effects.	✓	✓	✓	✓
tau p10/p50/p90	Distribution of estimated individual effects.	✓	✓	✓	✓
rc	Stata return code; nonzero means did not run.	✓	✓	✓	✓

One key distinction: PEHE evaluates heterogeneity directly, while ATE bias evaluates only the average effect.

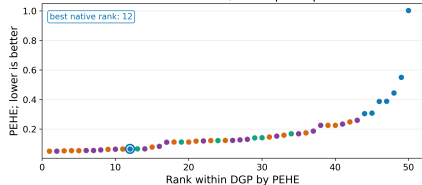
Simulation results: all estimators ranked by PEHE

Simulation PEHE rankings across all valid specifications

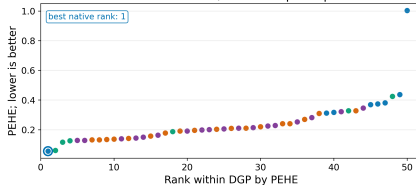
Four panels for the LL, LN, NL, and NN DGPs



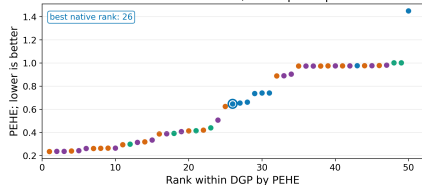
LL: linear outcome / linear participation



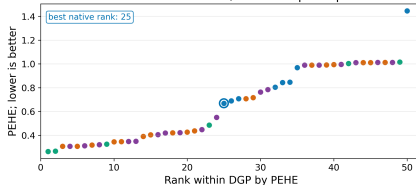
LN: linear outcome / nonlinear participation



NL: nonlinear outcome / linear participation



NN: nonlinear outcome / nonlinear participation



This figure uses every valid simulated result. The best methods differ sharply across DGPs: linear-friendly designs favor simple/orthogonal approaches, while nonlinear designs reward richer final CATE models.

Simulation: best eight specifications by DGP, Linear outcome

DGP	family	est	spec	xfit	ATE	bias	PEHE	corr	tau sd
LL	manual orth	irm	irm param reg-base	noxfit	1.005	0.007	0.051	1.00	0.887
LL	manual orth	irm	irm param reg-base	xfold5	1.001	0.003	0.051	1.00	0.882
LL	manual orth	irm	irm cv-lasso reg-base	noxfit	1.010	0.012	0.053	1.00	0.877
LL	manual orth	plr	plr param reg-base	noxfit	1.010	0.012	0.054	1.00	0.864
LL	manual orth	plr	plr cv-lasso reg-base	noxfit	1.010	0.012	0.054	1.00	0.866
LL	manual orth	plr	plr cv-lasso reg-base	xfold5	1.009	0.011	0.055	1.00	0.865
LL	manual orth	plr	plr param reg-base	xfold5	1.013	0.015	0.056	1.00	0.865
LL	manual orth	irm	irm cv-lasso reg-base	xfold5	1.018	0.020	0.059	1.00	0.874
LN	native cate	aipw	param reg/logit/reg	xfold5	0.982	-0.027	0.055	1.00	0.912
LN	non-DDML	plain	S-OLS base	noxfit	0.998	-0.004	0.061	1.00	0.886
LN	non-DDML	plain	t cv-lasso	noxfit	1.025	0.023	0.116	0.99	0.868
LN	non-DDML	plain	T-OLS rich	noxfit	1.010	0.008	0.125	0.99	0.891
LN	manual orth	plr	plr param reg-base	xfold5	0.968	-0.003	0.128	0.99	0.927
LN	manual orth	plr	plr cv-lasso reg-base	xfold5	0.970	-0.002	0.128	0.99	0.940
LN	manual orth	plr	plr cv-lasso reg-base	noxfit	0.973	0.001	0.132	0.99	0.943
LN	manual orth	plr	plr param reg-base	noxfit	0.972	0.000	0.132	0.99	0.937

Manual orthogonal IRM/PLR specifications dominate the linear DGPs.

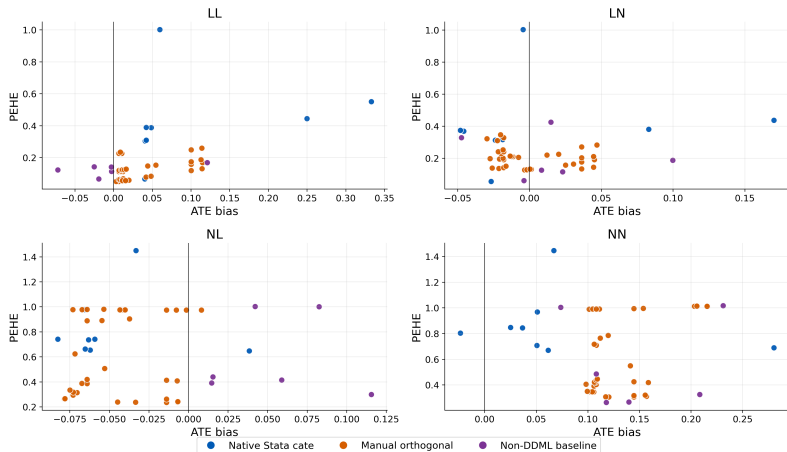
Simulation: best eight specifications by DGP, Non-linear outcome

DGP	family	est	spec	xfit	ATE	bias	PEHE	corr	tau sd
NL	manual orth	irm	irm plug-lasso reg-rich	noxfit	1.092	-0.014	0.235	0.98	1.031
NL	manual orth	irm	irm plug-lasso reg-rich	xfold5	1.099	-0.007	0.236	0.98	1.031
NL	manual orth	plr	plr plug-lasso reg-rich	xfold5	1.073	-0.033	0.237	0.98	1.092
NL	manual orth	plr	plr plug-lasso reg-rich	noxfit	1.061	-0.045	0.239	0.98	1.089
NL	manual orth	irm	irm plug-lasso cv-lasso	xfold5	1.099	-0.007	0.242	0.98	0.972
NL	manual orth	plr	plr cv-lasso reg-rich	xfold5	1.028	-0.078	0.261	0.97	1.045
NL	manual orth	irm	irm plug-lasso cv-lasso	noxfit	1.092	-0.014	0.262	0.98	0.932
NL	manual orth	plr	plr cv-lasso reg-rich	noxfit	1.028	-0.078	0.264	0.97	1.044
NN	non-DDML	plain	T-OLS rich	noxfit	1.206	0.118	0.264	0.98	1.012
NN	non-DDML	plain	t cv-lasso	noxfit	1.228	0.140	0.267	0.98	1.020
NN	manual orth	irm	irm plug-lasso cv-lasso	noxfit	1.201	0.145	0.307	0.97	1.061
NN	manual orth	plr	plr plug-lasso reg-rich	xfold5	1.177	0.120	0.307	0.97	1.097
NN	manual orth	plr	plr plug-lasso reg-rich	noxfit	1.174	0.117	0.307	0.97	1.092
NN	manual orth	irm	irm plug-lasso cv-lasso	xfold5	1.213	0.157	0.312	0.97	1.069
NN	manual orth	irm	irm plug-lasso reg-rich	noxfit	1.201	0.145	0.318	0.97	1.101
NN	manual orth	irm	irm plug-lasso reg-rich	xfold5	1.212	0.156	0.321	0.97	1.104

For nonlinear DGPs, rich final CATE models and lasso-based nuisance adjustment become more competitive.

Simulation: ATE bias is not PEHE

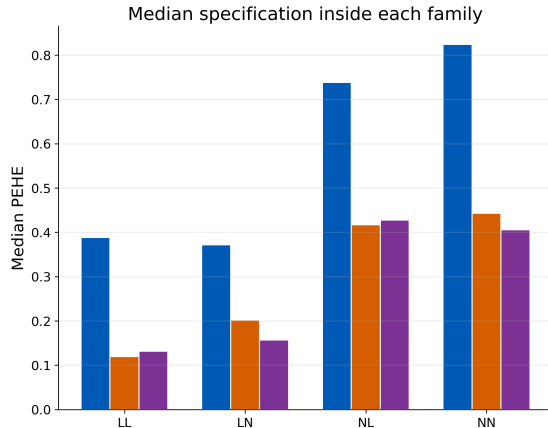
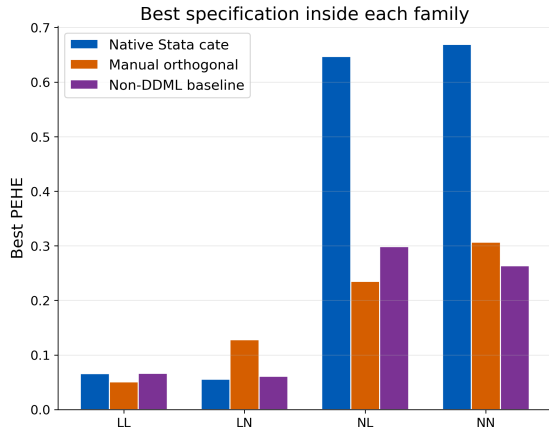
ATE accuracy versus CATE accuracy



A method can have small average-effect bias and still miss heterogeneity. This is the main reason to report CATE-specific metrics in addition to ATE metrics.

Simulation: median PEHE by estimator family

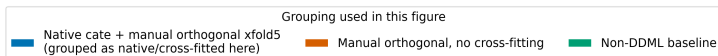
Do not confuse best-in-family performance with the median native-cate specification



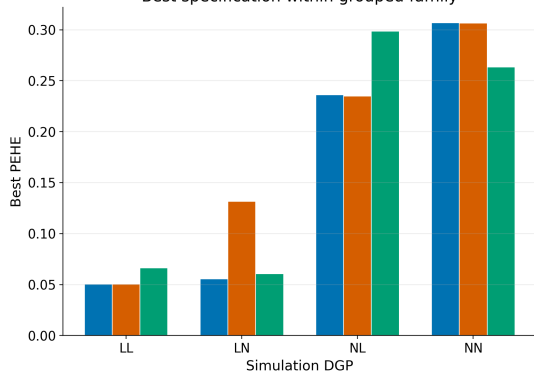
There is no single universal winner. The relative ranking changes when outcome surfaces and participation surfaces become nonlinear.

Simulation: median PEHE by estimator family

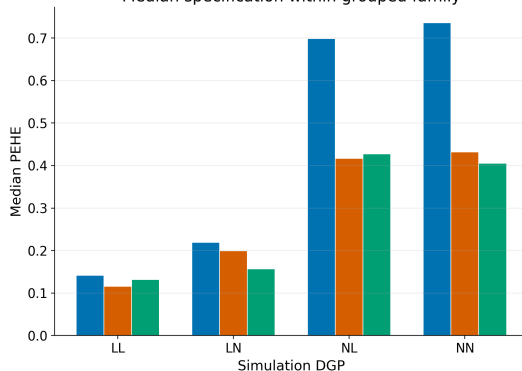
Simulation PEHE by grouped implementation family



Best specification within grouped family

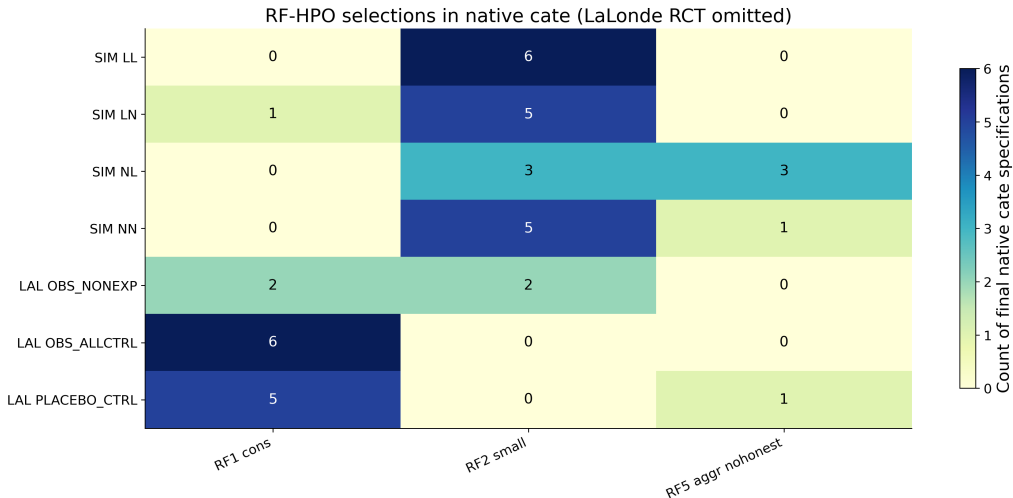


Median specification within grouped family



There is no single universal winner. The relative ranking changes when outcome surfaces and participation surfaces become nonlinear.

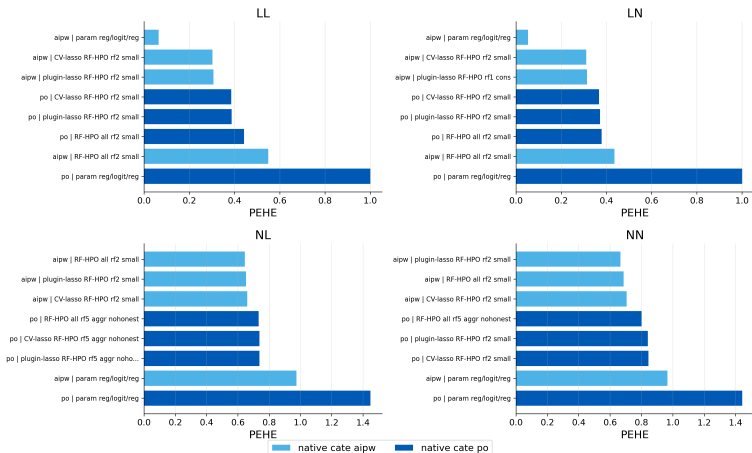
Simulation: RF-HPO changes across designs



HPO selects different RF candidates across DGPs and LaLonde designs. This makes RF parametrization an object of study rather than a harmless default.

Native CATE: HPO groups in the simulations

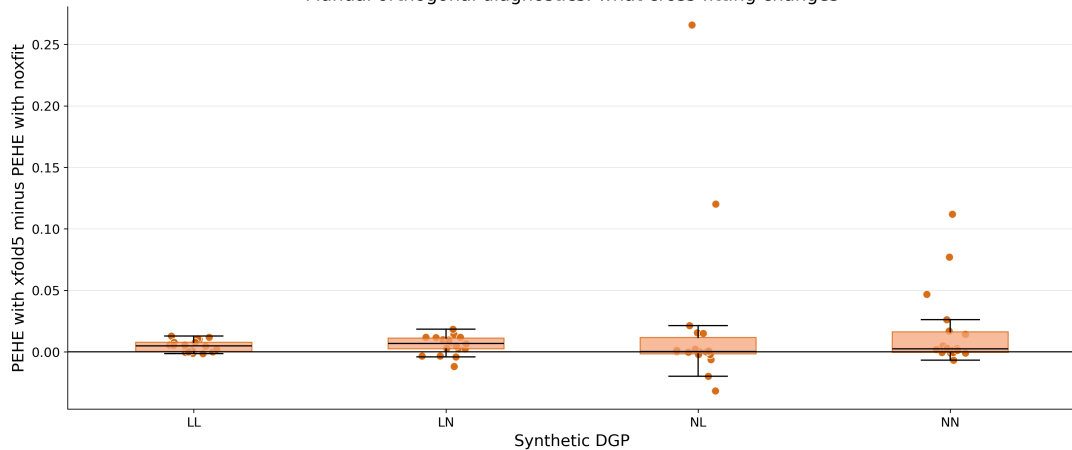
Native Stata cate only - performance differs across native specifications



In linear DGPs, parametric or simpler orthogonal methods often dominate. In nonlinear DGPs, RF-HPO improves over a purely parametric CATE model, but it does not automatically become best overall.

Manual orthogonal learners: effect of cross-fitting

Manual orthogonal diagnostics: what cross-fitting changes



Negative values mean cross-fitting lowered PEHE relative to the comparable no-cross-fit run.

Cross-fitting belongs to the DDML design. It is not HPO. In these finite samples, cross-fitting sometimes improves PEHE and sometimes slightly worsens it, which is why repeated splits may be worth exploring later.

Simulation: median performance by family and estimator

DGP	family	est	median PEHE	median —bias—	median corr	N
LL	manual orth	plr	0.103	0.015	0.99	12
LL	manual orth	irm	0.120	0.011	1.00	24
LL	non-DDML	plain	0.132	0.022	0.99	6
LL	native cate	aipw	0.307	0.042	0.94	4
LL	native cate	po	0.416	0.054	0.93	4
LN	non-DDML	plain	0.157	0.019	0.99	6
LN	manual orth	plr	0.185	0.012	0.99	12
LN	manual orth	irm	0.202	0.022	0.98	24
LN	native cate	aipw	0.314	0.025	0.94	4
LN	native cate	po	0.377	0.047	0.93	4
NL	manual orth	irm	0.417	0.059	0.95	24
NL	non-DDML	plain	0.427	0.051	0.93	6
NL	manual orth	plr	0.620	0.043	0.71	12
NL	native cate	aipw	0.658	0.064	0.81	4
NL	native cate	po	0.741	0.061	0.75	4
NN	non-DDML	plain	0.405	0.129	0.94	6
NN	manual orth	irm	0.442	0.110	0.93	24
NN	manual orth	plr	0.681	0.162	0.69	12
NN	native cate	aipw	0.698	0.056	0.76	4
NN	native cate	po	0.845	0.031	0.68	4

LaLonde/NSW-PSID designs

The local file contains the randomized NSW sample and a nonexperimental comparison group.

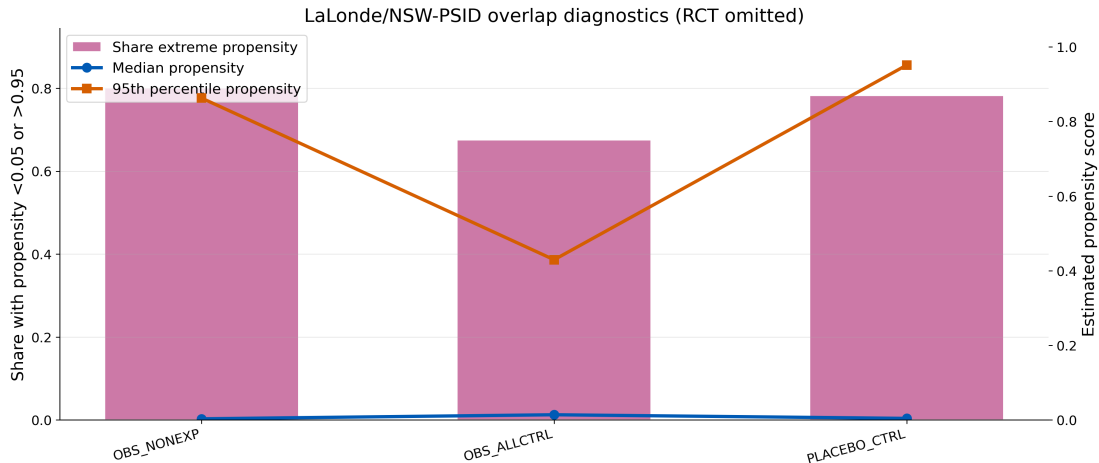
Design	Estimation sample	Purpose
RCT	Randomized treated and randomized controls.	Benchmark.
OBS-NONEXP	Randomized treated units vs nonexperimental PSID controls.	Classic observational stress test.
OBS-ALLCTRL	Randomized treated units vs randomized plus nonexperimental controls.	Checks whether adding controls changes estimates.
PLACEBO-CTRL	Randomized controls vs nonexperimental controls.	Selection/placebo diagnostic.

The RCT benchmark estimate in these results is about 886 dollars. The observational designs ask whether methods recover it without using randomized controls as the comparison group.

sample	N	treated	controls	bench	ps p1	ps p5	ps med	ps p95	ps p99	extreme
RCT	722	297	425	886	0.270	0.328	0.396	0.552	0.604	0.0%
OBS_NONEXP	2,787	297	2,490	886	0.000	0.000	0.003	0.863	0.965	80.1%
OBS_ALLCTRL	3,212	297	2,915	886	0.000	0.000	0.014	0.430	0.478	67.5%
PLACEBO_CTRL	2,915	425	2,490	0	0.000	0.000	0.004	0.951	0.977	78.3%

The observational and placebo designs have severe propensity-score overlap problems. This is the most important explanation for large benchmark failures.

LaLonde overlap: visual diagnostic

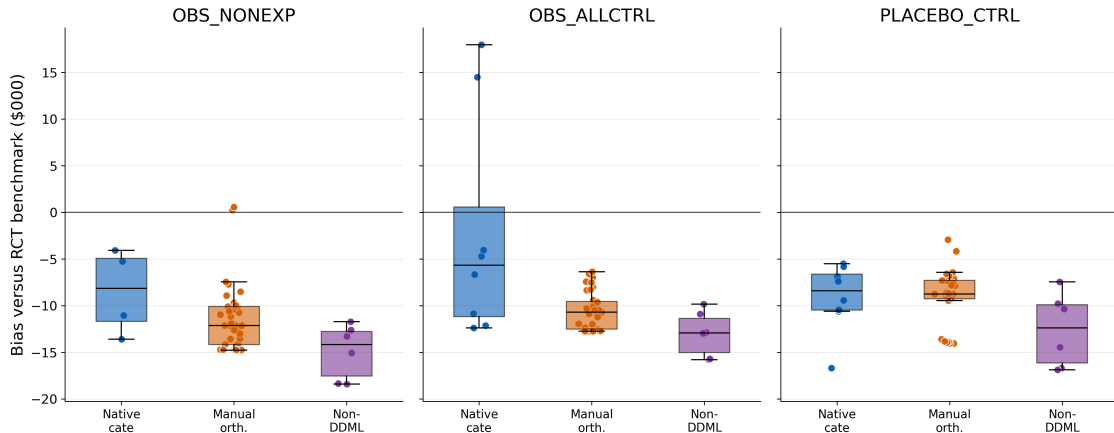


RCT is used as the benchmark only. The plotted rows use nonexperimental controls.

The RCT sample has broad overlap. The PSID-based comparisons concentrate many observations near 0 or 1, making causal ML much harder.

LaLonde: benchmark recovery for all valid specifications

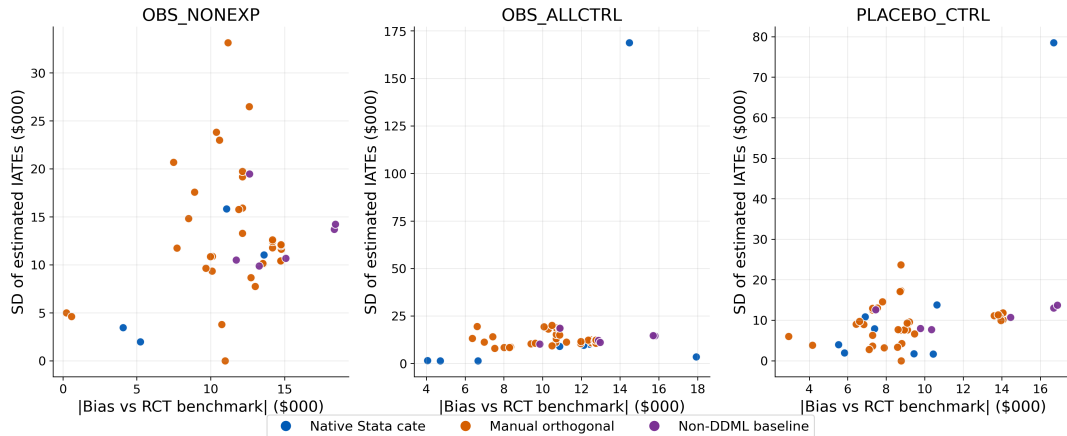
LaLonde benchmark recovery by design (RCT omitted from figure)



In the RCT sample, many methods are close to the benchmark. In observational samples, most methods miss the benchmark by thousands of dollars.

LaLonde: ATE benchmark match versus heterogeneity stability

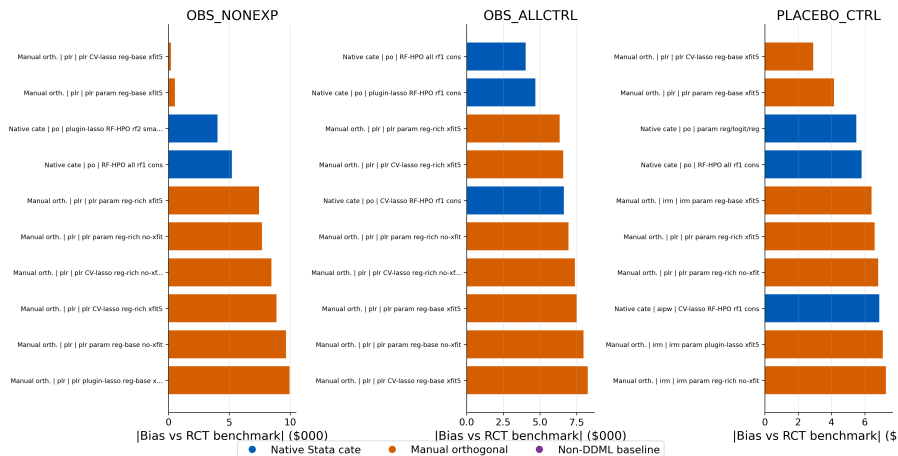
ATE benchmark fit and heterogeneity dispersion are different diagnostics



A method can match the benchmark while producing very dispersed individual effects. Both ATE recovery and CATE stability should be reported.

LaLonde: ten closest specifications to benchmark in each design

LaLonde: ten closest specifications to the RCT benchmark in each non-RCT design



The closest specifications differ by sample construction. In OBS-NONEXP, even the best specifications remain far from the RCT benchmark, which points to overlap/selection problems.

What the full results say so far

- **ATE and CATE are different targets.** Some specifications recover average effects but do poorly on PEHE or produce very dispersed individual effects.
- **Linear designs are forgiving.** In LL and LN, simple/manual orthogonal specifications often perform very well.
- **Nonlinear heterogeneity changes the ranking.** In NL and NN, rich final CATE models and flexible learners become more important.
- **HPO matters, but it is not magic.** RF-HPO can improve native cate results in nonlinear settings, but aggressive RF choices can also create instability.
- **The LaLonde observational samples are hard because of support.** The PSID comparisons have extreme propensities; most methods miss the RCT benchmark by large amounts.

What this means for applications

Do not report only the average effect.

- Report the ATE and the distribution of estimated CATEs.
- Check whether the best ATE estimator also gives credible heterogeneity.
- Examine support/overlap before interpreting subgroup or individual effects.
- Compare parametric, lasso, and RF-HPO versions, especially when there are nonlinearities or many controls.
- Treat HPO choices as part of the research design, not an implementation detail.

An appendix contains the detailed result tables:

- full native cate, non-DDML, and manual orthogonal simulation tables;
- full native cate, non-DDML, and manual orthogonal LaLonde/NSW-PSID tables;
- all slides previously numbered 27–42 and 49–70 in the full diagnostic deck.

References I



Bach, Philipp, Daniel Schacht, Victor Chernozhukov, and Stephan Klaassen. 2024. "Hyperparameter Tuning for Causal Inference with Double Machine Learning." *Proceedings of Machine Learning Research*.



Chernozhukov, Victor, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." *The Econometrics Journal* 21(1): C1–C68.



Knaus, Michael C., Michael Lechner, and Anthony Strittmatter. 2021. "Machine Learning Estimation of Heterogeneous Causal Effects: Empirical Monte Carlo Evidence." *The Econometrics Journal* 24(1): 134–161.



Künzel, Sören R., Jasjeet S. Sekhon, Peter J. Bickel, and Bin Yu. 2019. "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning." *Proceedings of the National Academy of Sciences* 116(10): 4156–4165.



Nie, Xinkun, and Stefan Wager. 2021. "Quasi-Oracle Estimation of Heterogeneous Treatment Effects." *Biometrika* 108(2): 299–319.



Wager, Stefan, and Susan Athey. 2018. "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests." *Journal of the American Statistical Association* 113(523): 1228–1242. doi: 10.1080/01621459.2017.1319839.