HIDDEN CONFOUNDING CAN MAKE CATE VALIDATION LOOK BETTER AS CAUSAL ERROR GROWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Because counterfactual outcomes are unavailable in observational data, conditional average treatment-effect (CATE) estimators are commonly developed and selected using observable held-out losses. A loss can rank candidates well without revealing whether any candidate remains causally accurate as the deployment environment changes. We uncover a systematic directional failure under hidden confounding: as confounding strengthens, causal error (oracle CATE root-PEHE) rises while held-out R-loss and a doubly robust pseudo-outcome loss improve. A population construction explains the direction: under hidden confounding, both losses reward movement toward the observed treated–untreated outcome difference rather than the CATE. Guided by this mechanism, we audit 24 structurally varied environments. Eight estimators spanning classical meta-learners, forests, orthogonal methods, and neural architectures exhibit this *validation reversal* in more than 92% of paired severity comparisons. Observable validation can still select a near-oracle-best member even as the best attainable CATE accuracy of the entire candidate pool deteriorates. We therefore propose severity-path validation audits as a new evaluation framework for causal-estimator development: developers should test whether observable model-selection criteria remain aligned with oracle causal accuracy as identification is progressively stressed.

1 INTRODUCTION

Causal effect estimation from observational data begins with a conditional-randomization argument. If an adjustment set X makes treatment independent of the potential outcomes, $(Y(0), Y(1)) \perp T \mid X$ (conditional unconfoundedness), and consistency and positivity hold, treated and untreated units are conditionally exchangeable. Within levels of X they may be compared as if treatment were randomized, identifying $\mathbb{E}[Y(t) \mid X = x] = \mathbb{E}[Y \mid T = t, X = x]$ (Rubin, 1974; Pearl, 2009; Rosenbaum & Rubin, 1983). Neyman's potential-outcome formulation provides the underlying notation (Splawa-Neyman, 1990). The crucial difficulty is that observed data alone generally cannot verify that X is sufficient: an omitted common cause can invalidate the causal interpretation while leaving the observed distribution apparently well behaved.

Estimation and evaluation are coupled stages of causal analysis. *Estimation* constructs candidate conditional treatment-effect (CATE) functions—an estimator *pool*. *Evaluation* uses held-out data to choose a candidate and assess its performance. Practitioners use evaluation to guide estimator revision, even when identifying assumptions cannot be verified from observed data. We study two widely used families of observable CATE criteria. R-loss residualizes both outcome and treatment and scores the candidate through their orthogonalized product; DR loss instead constructs an augmented inverse- propensity pseudo-outcome and applies squared error to the candidate (Schuler et al., 2018; Saito & Yasui, 2020; Nie & Wager, 2021; Mahajan et al., 2024). Their nuisance structures and finite-sample behavior differ, making them complementary views of an estimated CATE: R-loss evaluates an orthogonalized residual relation, whereas DR loss evaluates squared error against an augmented inverse-propensity pseudo-outcome. Both are computable from observed held-out data and therefore can guide model development when counterfactual error itself is unavailable. Our question is whether these two observable criteria continue to track causal accuracy as identification weakens.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Much CATE validation asks which candidate an observable score should select on benchmark panels (Alaa & van der Schaar, 2019; Saito & Yasui, 2020; Mahajan et al., 2024); controlled studies also vary confounding, misspecification, or imbalance (Curth & van der Schaar, 2023). Econometric sensitivity analysis asks how causal conclusions change under specified departures from identifying assumptions (Rosenbaum, 2002; Cinelli & Hazlett, 2020), and related machine-learning work constructs uncertainty estimates, bounds, or deferral rules under overlap and confounding stress (Jesson et al., 2020; 2021). We bring the evaluation criterion itself into this analysis. Causal systems routinely move from development to deployment across populations, sites, and time periods whose treatment-assignment mechanisms differ. We therefore ask: *as hidden confounding strengthens across a controlled family of deployment environments, do held-out observational losses remain directionally aligned with causal accuracy—or can they improve while the estimator that they rank best becomes less causally accurate?* To answer, we implement a paired latent-confounding stress path starting from a valid severity-zero baseline and follow score direction, absolute causal error, and relative selection along that path. At successive severity levels, each seed reuses its covariates, latent variables, potential outcomes, and outcome noise while only the treatment mechanism changes; pairing therefore isolates stronger treatment dependence on an unobserved cause of outcome U from unrelated sampling variation. We instantiate additive and nonlinear latent outcome contributions while retaining broad overlap in the observed-X propensity. This extends robustness analysis from the causal conclusion itself to the observable evidence used to develop and select the estimator. A separate *proxy study* applies the same hidden-confounding stress while making U indirectly observable through three pretreatment measurements $Z_k = U + \epsilon_k$; distinct conditions increase their noise or replace them with uninformative measurements. Consistency and absence of interference hold throughout.

We test two broad families of CATE estimators. The first spans four widely used classical and meta-learning implementations from EconML (Battocchi et al., 2019): XLearner (Künzel et al., 2019), DRLearner (Kennedy, 2023), CausalForestDML (Wager & Athey, 2018; Athey et al., 2019), and NonParamDML (Chernozhukov et al., 2018). The second spans four neural architectures: TARNet and CFRNet (Shalit et al., 2017), DragonNet (Shi et al., 2019), and FlexTENet (Curth & van der Schaar, 2021), using their authors' released implementations. Together, these eight estimators cover distinct learning objectives, nuisance-adjustment strategies, and model classes, allowing us to test whether validation reversal transfers across substantially different approaches to CATE estimation. In the base and environment-bank experiments, training, selection, monitoring, and oracle-test samples are mutually independent within each cell, while each split's underlying random draws are paired across severity. First, every candidate is fitted using only the training sample. We then evaluate every fixed candidate with both R-loss and DR loss on an independent monitoring sample. Following each fixed estimator identity—refitted at every severity—establishes that reversal does not depend on switching among estimators. For relative model selection, selection-split R-loss chooses one EconML pool member, and we report both monitoring losses and oracle error for that choice; this measures how relative selection quality coexists with changing absolute accuracy. The oracle-test sample measures error against the true CATE. Both pools rely on identification from observed covariates. We examine a particularly consequential response to weakening these assumptions: causal error rises while an independently held-out validation loss falls and appears to improve. We call this directional pattern a *validation reversal*. Selection regret separately measures how closely the chosen candidate approaches the best oracle error available in its pool.

We next examine methods that receive partial information about a threatened assumption. CEVAE (Louizos et al., 2017) models a latent confounder from observed proxies. CausalPFN (Balazadeh Meresht et al., 2025) amortizes CATE estimation over synthetic tasks satisfying ignorability and returns a predictive interval. We test CEVAE's accuracy as proxy information weakens and examine whether CausalPFN's model-based uncertainty reflects the accompanying causal error.

We find systematic reversal under hidden confounding. Figure 1 shows the central pattern in the additive and nonlinear base designs: oracle root-PEHE rises with confounding severity while held-out R-loss falls. Across 24 structurally varied environments, the paired event "root-PEHE rises while held-out loss falls" occurs in 92–100% of environment–seed endpoint comparisons, depending on estimator and score. R-loss reverses in 93–100% of paired comparisons, and DR loss reverses in 92–97%. The result spans all four EconML estimators and TARNet, CFRNet, DragonNet, and FlexTENet. Fixed-candidate trajectories establish the reversal without selection, while the model-selection analysis shows that R-loss can still choose a near-oracle-best pool member as the pool's absolute causal accuracy deteriorates. The proxy study shows the same tension within methods given

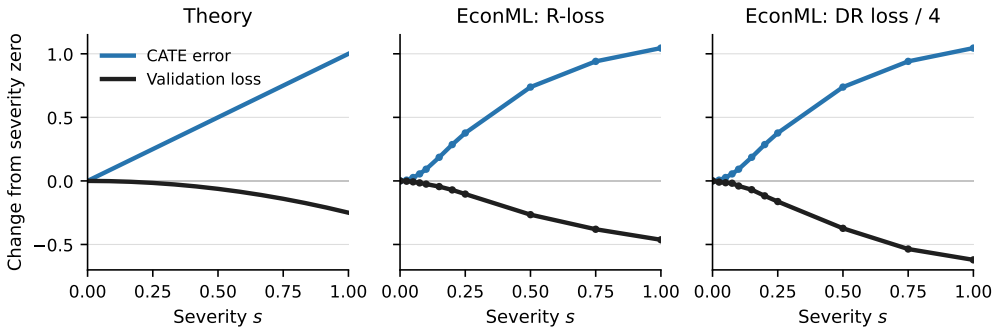


Figure 1: **Theory and experiment exhibit the same validation reversal.** All panels report change from severity zero on identical axes. Left: the population construction of Section 3, in an illustrative parameterization where CATE root-mean-squared error is s and R-loss changes by $-s^2/4$. Middle and right: empirical means across the additive and nonlinear hidden-confounding families. Selection-split R-loss chooses among the four EconML candidates; blue shows the chosen candidate's oracle-test root-PEHE, while black shows its monitoring R-loss or DR loss. DR loss is divided by four for a common visual scale; this rescaling does not change its direction. Each empirical curve averages 30 paired seeds within each family and then averages the two family means; the band is their pointwise range.

partial information about the confounder. As the proxies become uninformative, fixed CEVAE's severity-zero-to-one root-PEHE increase grows from .106 to .900, while its independently monitored R-loss changes from nearly flat (+.005) to a .454 decrease; CausalPFN's interval width becomes increasingly negatively associated with causal error. For developers of causal estimators and validation criteria, the practical implication is to audit score trajectories under graded hidden-confounding stress, not candidate rankings alone.

We contribute a severity-path validation audit that pairs graded hidden-confounding stress with independent monitoring and three trajectories: selected causal error, the pool's oracle-best error, and the observable score. Within this framework, we provide (i) a precise separation of absolute causal error, the best error available within an estimator pool, and the quality of observable-score selection; (ii) a theoretical population construction that identifies the common mechanism (Section 3): as hidden confounding increases, both validation losses fall while CATE error rises, and within each confounded population both losses strictly prefer the biased observational target over the true CATE; and (iii) an operationalization of this theory as paired stress tests of R-loss and DR loss across eight effect estimators and a balanced bank of 24 environments, supplemented by a proxy-information study of CEVAE and CausalPFN. Together they distinguish better relative selection from worsening absolute causal accuracy—even when both accompany an improving validation score.

2 Related work and positioning

CATE validation and selection. Influence-function criteria, factual prediction, nearest-neighbor scores, pseudo-outcome losses, and orthogonal objectives have all been proposed for selecting treatment-effect estimators without counterfactual labels (Alaa & van der Schaar, 2019; Schuler et al., 2018; Nie & Wager, 2021). Recent broad comparisons emphasize that performance depends on the candidate pool, response surface, and nuisance estimation, and caution against a universal winner (Curth & van der Schaar, 2023; Mahajan et al., 2024). This literature evaluates a surrogate mainly by whether it ranks candidates with low oracle error. Our outcome differs: whether the numerical score improves while selected absolute risk worsens along a paired severity path. Low regret and score–risk reversal can therefore coexist.

Benchmarks and stress tests. Semi-synthetic benchmarks provide potential-outcome ground truth but commonly compare estimators at a fixed data-generating process (Hill, 2011; Dorie et al., 2019). IncomeSCM shows that models with similar factual fit can yield very different causal quality (Johansson, 2024), while validation studies examine whether causal assumptions are plausible in a

given dataset (Parikh et al., 2022). We complement these approaches with paired severity paths that vary hidden-confounding strength. Holding exogenous draws and response surfaces fixed reduces variation unrelated to the challenged assumption and makes score direction interpretable rather than merely comparing average ranks over unrelated datasets.

Sensitivity analysis and identification. Sensitivity analysis, negative controls, and proxy methods add explicit structure or information (Rosenbaum, 2002; Cinelli & Hazlett, 2020; Lipsitch et al., 2010; Miao et al., 2018). Sensitivity analysis asks how conclusions change over specified departures from identification. Our analysis follows the observed selection score itself, comparing its trajectory with absolute causal error on an independent monitoring split. The population R-loss calculation below explains how the observed-data objective can favor an associational target as identifying assumptions weaken.

Neural, latent, and pretrained estimators. TARNet and CFRNet use shared representations, with CFRNet additionally balancing treated and control representations (Shalit et al., 2017); DragonNet jointly models treatment and outcomes with targeted regularization (Shi et al., 2019); CEVAE uses proxies in a latent-variable model (Louizos et al., 2017); and CausalPFN amortizes effect estimation in a pretrained transformer (Balazadeh Meresht et al., 2025). These methods span very different inductive biases, but validity remains conditional on supplied information and assumptions. We test whether architectural sophistication or model uncertainty provides an implicit safeguard.

3 Why observable validation can reward confounded effects

This section constructs a single severity-indexed family in which hidden confounding strengthens, the effect remains heterogeneous in observed covariates, and two population validation losses move opposite to CATE accuracy. The construction isolates the mechanism at the population optimum; the empirical study measures how broadly the same directional relationship appears across finite-sample learners. Figure 1 places this population trajectory beside its empirical counterparts.

Let X be arbitrary observed covariates and let $U \in \{-1, 1\}$ be independent of X, with equal probability on each value. For square-integrable functions b, a, and τ, with $A := \mathbb{E}[a(X)^2] > 0$, define

$$Y(t) = b(X) + a(X)U + t\tau(X) + \epsilon, \qquad \Pr_s(T = 1 \mid X, U) = \frac{1 + sU}{2}, \quad 0 \le s \le 1, \quad (1)$$

where $\epsilon \perp (X, U, T)$, $\mathbb{E}[\epsilon] = 0$, and $\mathrm{Var}(\epsilon) = \sigma^2$. We assume consistency, so $Y = Y(T)$. The causal target is $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$. For binary U, the baseline form $b(X) + a(X)U$ is fully general for the conditional mean of $Y(0)$ given (X, U): any function $q_0(x, u)$ can be represented by choosing $b(x) = \{q_0(x, 1) + q_0(x, -1)\}/2$ and $a(x) = \{q_0(x, 1) - q_0(x, -1)\}/2$. Within this representation, we deliberately restrict $Y(1) - Y(0) = \tau(X)$ to exclude additional U-dependent effect modification. We also hold P_X fixed across severity, take $U \perp X$ to be equiprobable, and use an assignment mechanism with no direct X dependence; the latter assumptions keep the observed propensity fixed. Together these choices hold the causal response surface and CATE fixed, so reversal is driven by latent selection rather than a changing causal estimand. Severity changes only the dependence of treatment on the outcome-relevant U: treatment is conditionally unconfounded at $s = 0$, while hidden confounding strengthens with s (indeed, $\mathrm{Corr}_s(T, U \mid X) = s$). Meanwhile $e_s(x) := \Pr_s(T = 1 \mid X = x) = 1/2$ for every x and s, so observed-covariate overlap is fixed.

Using the exact observed-data nuisances, let $m_s(x) = \mathbb{E}_s[Y \mid X = x] = b(x) + \tau(x)/2$. Robinson's transformation (Robinson, 1988) defines population R-loss (Nie & Wager, 2021)

$$L_{R,s}(g) = \mathbb{E}_s[(Y - m_s(X) - (T - e_s(X))g(X))^2], \qquad g_{R,s} := \arg\min_{g \in L^2(P_X)} L_{R,s}(g). \quad (2)$$

Conditional least squares and binary treatment give

$$\begin{aligned} g_{R,s}(x) &= \mathbb{E}_s[Y \mid T = 1, X = x] - \mathbb{E}_s[Y \mid T = 0, X = x] \\ &= \tau(x) + 2s\, a(x). \end{aligned} \quad (3)$$

Thus the population target equals the CATE at $s = 0$ and moves away from it in the direction of the latent outcome contribution as confounding increases; $2s\, a(X)$ is the induced hidden-selection bias.

The standard DR pseudo-outcome provides a complementary route to the same observational target. Writing $\mu_{t,s}(x) = \mathbb{E}_s[Y \mid T = t, X = x]$, define

$$\phi_{DR,s} = \mu_{1,s}(X) - \mu_{0,s}(X) + \frac{T\{Y - \mu_{1,s}(X)\}}{e_s(X)} - \frac{(1-T)\{Y - \mu_{0,s}(X)\}}{1 - e_s(X)}, \tag{4}$$
$$L_{DR,s}(g) = \mathbb{E}_s[(\phi_{DR,s} - g(X))^2].$$

This pseudo-outcome underlies widely used CATE validation criteria (Saito & Yasui, 2020; Schuler et al., 2018). Exact observed-data nuisances imply $\mathbb{E}_s[\phi_{DR,s} \mid X = x] = \mu_{1,s}(x) - \mu_{0,s}(x) = g_{R,s}(x)$, so both losses have the same confounded population minimizer. Suppose the estimator is refitted at each severity and attains this common optimum: $g_s = g_{R,s}$.

Proposition 1 (Validation reversal along increasing hidden confounding)**.** *Under Equation 1, the squared CATE error of g_s strictly increases with s, while its population R-loss and DR loss strictly decrease:*

$$\mathbb{E}[(g_s(X) - \tau(X))^2] = 4s^2A,$$
$$L_{R,s}(g_s) = \sigma^2 + (1 - s^2)A, \qquad L_{DR,s}(g_s) = 4\{\sigma^2 + (1 - s^2)A\}. \tag{5}$$

Proof. The assignment mechanism gives $\mathbb{E}_s[U \mid T = 1, X] = s$ and $\mathbb{E}_s[U \mid T = 0, X] = -s$, proving $g_s = \tau + 2sa$ and the CATE-error formula. Let $W = 2T - 1$. Because $\mathbb{E}_s[UW \mid X] = s$, the R-loss residual at g_s is $a(X)\{U - sW\} + \epsilon$, whose squared expectation is $\sigma^2 + (1 - s^2)A$. The exact arm means are $\mu_{1,s} = b + \tau + sa$ and $\mu_{0,s} = b - sa$; substituting them into Equation 4 gives $\phi_{DR,s} = \tau + 2W\{a(X)U + \epsilon\}$ and therefore the stated DR loss. Since $A > 0$, the claimed monotonicities follow. □

Corollary 2 (Observable validation rewards confounding-induced bias)**.** *The true CATE has severity-invariant losses $L_{R,s}(\tau) = \sigma^2 + A$ and $L_{DR,s}(\tau) = 4(\sigma^2 + A)$. At every fixed severity $s > 0$, both criteria therefore assign strictly lower loss to the causally biased $g_s = \tau + 2sa$ than to the true CATE:*

$$L_{R,s}(\tau) - L_{R,s}(g_s) = s^2A, \qquad L_{DR,s}(\tau) - L_{DR,s}(g_s) = 4s^2A. \tag{6}$$

Thus the validation advantage of the confounding-induced bias grows quadratically with severity.

Proof. Substitute $g = \tau$ into the two population losses and subtract the values in Equation 5. Positivity of A makes both gaps strict for $s > 0$. □

Because $e_s \equiv 1/2$ and $m_s = (\mu_{1,s} + \mu_{0,s})/2$, the pointwise identity $\phi_{DR,s} - g = 2(2T - 1)\{Y - m_s - (T - 1/2)g\}$ implies $L_{DR,s}(g) = 4L_{R,s}(g)$ for every candidate function g. This proportionality follows from balanced assignment and matched exact nuisances, rather than from the outcome model. The empirical study relaxes both conditions and asks how often the shared directional mechanism appears for finite-sample learners as hidden confounding increases.

4 Experimental design

4.1 Data-generating mechanisms

The empirical design translates the mechanism isolated in Section 3 into a finite-sample stress test. As in Equation 1, outcomes contain an observed baseline, an outcome-relevant latent contribution, and a heterogeneous treatment effect, while severity changes only the latent variable's influence on treatment assignment. We retain that structure but replace binary U and balanced assignment with continuous U, nonlinear response functions, and an X-dependent logistic propensity. The potential-outcome model below is therefore the empirical analogue of Equation 1, and Equation 7 is its more realistic severity-indexed assignment mechanism.

The study has two tiers. Two transparent base families—additive and nonlinear in the latent outcome contribution—translate Section 3 into a finite-sample setting with continuous U, X-dependent assignment, estimated nuisances, and fitted learners. They make the mechanism inspectable and establish the paired-split evaluation protocol. The base studies therefore ask whether the theoretical mechanism appears in explicit, realistic instances; they are not the breadth claim. A 24-environment bank, specified before its outcomes were inspected, then asks whether the result transfers beyond

those equations. The bank is a balanced fractional design: each covariate law appears four times, each response, CATE, and latent category six times, and each treatment form eight times. The same latent-assignment stress and fixed estimator settings are used throughout. Appendix A records the protocol, and Appendix D lists every structural combination. We draw $X \in \mathbb{R}^{10}$ standard Gaussian. Let $Y(0)$ and $Y(1)$ denote the untreated and treated potential outcomes. Their conditional means, given X and U, are $\mu_0(X,U) = b(X) + h(U)$ and $\mu_1(X,U) = \mu_0(X,U) + \tau(X)$. We generate $Y(t) = \mu_t(X,U) + \epsilon$, where the same $\epsilon \sim N(0,1)$ is independent of (X,U,T) for both potential outcomes, and observe only $Y = Y(T)$. Thus $b(X)$ describes dependence on observed covariates, $h(U)$ is any latent contribution, and $\tau(X) = \mu_1 - \mu_0$ is the true treatment effect. Specifically, $b(X) = .7X_1 + .5\sin X_2 + .25X_3^2 - .2X_4X_5$ and $\tau(X) = 1 + .6\tanh X_1 + .35X_2 - .2X_3X_4$. For hidden confounding, $U \sim N(0,1)$ and treatment assignment follows

$$\pi_s(X,U) = \Pr(T = 1 \mid X, U) = \text{logit}^{-1}(.35X_1 - .25X_2 + 3sU). \tag{7}$$

Here $h(U) = U$ in the additive family and $h(U) = U + .35(U^2 - 1)$ in the nonlinear family. Severity changes U's influence on assignment while $h(U)$ remains fixed: at $s = 0$, U affects outcomes and is independent of treatment given X; as s grows, the same U increasingly sorts units into treatment groups. The observed-data propensity is instead $e_s(X) = \Pr(T = 1 \mid X) = \mathbb{E}[\pi_s(X,U) \mid X]$. It retains broad support throughout the path, so this experiment isolates the consequences of omitting the outcome-relevant U while preserving broad observed-X overlap. The proxy study uses the same latent treatment mechanism but adds three pretreatment measurements $Z_k = U + \epsilon_k$, $\epsilon_k \sim N(0, \sigma^2)$, with $\sigma \in \{.5, 1, 2\}$ in separate conditions; another condition replaces them with independent-noise measurements. Proxy quality changes the information about U while outcome and treatment mechanisms remain fixed. This separates latent confounding from observed-covariate overlap concerns (D'Amour et al., 2021). Figure 2 shows the causal arrow controlled by severity.

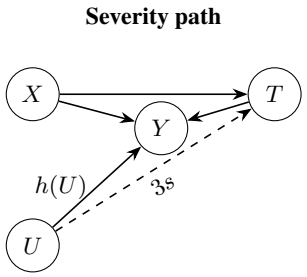


Figure 2: Hidden-confounding stress path. The dashed arrow marks the assignment dependence strengthened by s; the outcome mechanism and all solid arrows remain fixed. At $s = 0$, $U \to Y$ is present and $U \to T$ has coefficient zero, making U and T conditionally independent given X. The proxy experiment additionally supplies measurements Z_k of U.

4.2 Evaluation protocol

An experimental *cell* comprises training and evaluation samples at a fixed family or environment, severity, seed, and proxy condition where applicable. A seed reuses the same covariates, latent variables, potential outcomes, outcome noise, and treatment-assignment uniform draws across severities; treatment is coupled as $T_s = \mathbf{1}\{V < \pi_s(X,U)\}$ with shared V. Each fitted candidate is scored on the same independent monitoring sample by R-loss (Equation 2), which uses propensity e and outcome regression m, and DR loss (Equation 4), which uses e and arm-specific regressions μ_0, μ_1 to form a pseudo-outcome. These nuisances are fitted on training data only. Neither score receives U, potential outcomes, or the true CATE; lower values conventionally indicate better estimates. This implements the two validation routes analyzed together in Section 3.

Oracle evaluation is kept separate. For candidate f at severity s, root-PEHE is $\mathcal{R}_s(f) = \{\mathbb{E}_s[(\hat{\tau}_f(X) - \tau(X))^2]\}^{1/2}$ on the oracle-test sample. The primary audit follows each fixed candidate across severity and asks whether either monitoring loss falls while $\mathcal{R}_s(f)$ rises. Fixed identities separate score behavior from changes in the winner. Across the environment bank, this audit covers four EconML and four neural methods—TARNet, CFRNet, DragonNet, and FlexTENet—under both losses.

We separately examine relative selection. Selection-split R-loss chooses $\hat{f}_s = \arg\min_f \mathcal{V}_s^R(f)$ within the EconML pool; we then report its monitoring losses, selected risk $\mathcal{R}_s(\hat{f}_s)$, oracle-best risk $\min_f \mathcal{R}_s(f)$, and their difference, selection regret. This decomposition tests whether a score can choose competitively even as the pool becomes less causally accurate. Paired changes from severity zero are primary. Appendix E reports the corresponding base-family neural selection analysis and its exact pool.

The base studies use 2,000/1,000/2,000/5,000 training, selection, monitoring, and oracle-test observations, 11 severities, and 30 paired seeds. The bank uses 24 environments, 11 severities, and 20 paired seeds. Appendix A gives complete grids, sample sizes, software, nuisances, and fitting settings.

Targeted analyses ask whether additional information changes this picture. In the proxy study, we fit CEVAE (Louizos et al., 2017), CausalPFN (Balazadeh Meresht et al., 2025), DRLearner, and causal forest using only $[X, Z]$; two estimators given true U provide finite-sample reference trajectories. We inspect fixed-estimator curves and CausalPFN's reported interval width. The proxy study selects on its validation sample and recomputes R-loss on an independently seeded monitoring sample used neither for fitting nor selection; causal error is measured on an oracle-test split. Its four proxy conditions yield 1,320 cells.

To determine whether the R-loss reversal is caused by imperfect nuisance estimation rather than the criterion's observational target, we hold every fitted effect estimate fixed and recompute its score with two stronger nuisance specifications: flexible cross-fitting and the exact observed-X outcome and propensity functions. Persistence under these replacements isolates score–target alignment from nuisance-fitting error (Appendix C).

5 Results

We use $\mathcal{R}_s(\hat{f}_s)$ for the oracle root-PEHE of the score-selected estimator at severity s, following the definition at the start of Section 4.2. The oracle-best risk $\min_f \mathcal{R}_s(f)$ and their difference (selection regret) are always computed within the same pool. Results cover four EconML estimators and four neural estimators: TARNet, CFRNet, DragonNet, and FlexTENet.

5.1 Hidden confounding raises causal error while observable validation improves

We begin with the two explicit hidden-confounding DGPs defined in Section 4.1: the additive and nonlinear latent-outcome families. These base designs provide a direct empirical counterpart to the mechanism in Section 3; the broader 24-environment transfer study follows in Section 5.2. Figure 1 shows the EconML pool's paired severity trajectories in the two base DGPs. In both, the selected estimator's causal error rises while its independently measured R-loss falls. Appendix B gives the unnormalized endpoint values, paired intervals, and selected-versus-oracle-best decomposition for the EconML pool and the neural estimators.

Table 1: **Base-family reversal for the R-loss-selected EconML estimator.** Severity-one minus severity-zero changes over 30 paired seeds; both losses use independent monitoring data.

Family	Root-PEHE $0 \to 1$	Δ root-PEHE	Δ R-loss	Δ DR loss	Regret_1
Additive	$.374 \to 1.411$	$+1.036$	$-.458$	-2.482	.006
Nonlinear	$.368 \to 1.422$	$+1.054$	$-.473$	-2.486	.013

Table 1 shows that selected risk nearly quadruples, but endpoint regret remains $.006$–$.013$: oracle-best risk itself rises by about 1.05, so R-loss selects close to the best member even as the pool's best attainable risk deteriorates. Monitoring-split R-loss falls from 2.235 to 1.777 (additive) and 2.504 to 2.031 (nonlinear): the held-out criterion improves precisely as causal accuracy deteriorates. Independently monitored DR loss, computed for the same R-loss-selected candidates with separately fitted arm-specific nuisances, falls by 2.482 (95% paired interval $[-2.648, -2.311]$) and 2.486 ($[-2.624, -2.351]$), respectively. Thus the base reversal is already shared by both validation families before the broader transfer experiment.

Two controls test the roles of nuisance estimation and score scale. First, five-fold cross-fitted estimates and exact observed-X functions $m(X) = \mathbb{E}[Y \mid X]$, $e(X) = \Pr(T = 1 \mid X)$ preserve the inverse score–error relationship while effect estimates stay fixed. Second, we subtract the same-severity R-loss of the true CATE and normalize this excess by the gap between the zero-effect and true-CATE scores. The relationship persists in both controls. These simulation comparisons locate the reversal beyond nuisance error or an additive shift in loss level; Appendix C gives formulas and continuous score–error correlations.

The endpoint decomposition separates pool accuracy from selection quality. Under hidden confounding, R-loss selects close to the oracle-best member, yet even that member's root-PEHE rises by about 1.0 as severity rises. The improving score therefore misrepresents the pool's absolute causal accuracy while still choosing a near-best member. Comparing selected risk, oracle-best risk, and the score reveals whether evaluation fails as an absolute measure, as a selector, or both.

5.2 Validation reversal transfers across 24 environments

The base families establish the reversal in two explicit equations. We next test its breadth across structural variation and with DR loss, whose arm-specific outcome regressions and inverse-propensity augmentation differ from Robinson residualization. Under the construction of Section 3, the two criteria share a population target; the experiment tests whether that agreement persists with estimated nuisances, nonconstant propensities, and finite-sample learners. We use a balanced bank of 24 environments spanning six covariate laws, four response surfaces, four CATE forms, one- and three-dimensional latent variables, and three treatment-assignment forms. Each environment contains 11 severities and 20 paired seeds. Table 2 summarizes endpoint magnitudes; Appendix Table 7 reports estimator-wise reversal prevalence, and Appendix Table 6 lists every structural combination.

Table 2: **Magnitude of validation reversal across 24 environments.** Entries are medians across environment-level mean severity-zero-to-one changes, with interquartile ranges in brackets. Positive Δroot-PEHE means worsening causal accuracy; negative loss changes mean apparently improving validation. "Selected" denotes R-loss selection within the stated pool.

Family	Candidate	Δ root-PEHE	Δ R-loss	Δ DR loss
EconML	XLearner	$+1.134$ $[.659, 1.859]$	$-.599$ $[-1.248, -.176]$	-4.396 $[-7.446, -1.439]$
	DRLearner	$+.889$ $[.457, 1.584]$	$-.623$ $[-1.434, -.184]$	-4.397 $[-7.613, -1.441]$
	CausalForestDML	$+1.140$ $[.694, 1.950]$	$-.596$ $[-1.249, -.172]$	-4.382 $[-7.432, -1.447]$
	NonParamDML	$+.938$ $[.537, 1.596]$	$-.608$ $[-1.278, -.190]$	-4.419 $[-7.583, -1.439]$
	Selected (four)	$+1.111$ $[.690, 1.941]$	$-.509$ $[-1.250, -.173]$	-4.414 $[-7.426, -1.412]$
Neural	TARNet	$+.990$ $[.639, 1.848]$	$-.595$ $[-1.251, -.178]$	-4.189 $[-6.972, -1.260]$
	CFRNet	$+.977$ $[.565, 1.806]$	$-.587$ $[-1.219, -.179]$	-3.819 $[-6.967, -1.202]$
	DragonNet	$+1.004$ $[.642, 1.910]$	$-.620$ $[-1.425, -.176]$	-4.188 $[-6.741, -1.289]$
	FlexTENet	$+.912$ $[.485, 1.608]$	$-.544$ $[-1.269, -.164]$	-4.092 $[-7.099, -1.307]$

The magnitude is consistent across estimator summaries. In the typical environment, root-PEHE rises by about .89–1.14, R-loss falls by .51–.62, and DR loss falls by 3.82–4.42. Table 2 reports the distribution of environment-level mean changes through medians and interquartile ranges. Environment-clustered means and intervals are included in the supplementary results. The direction is stable: across the EconML candidates, R-loss reverses in 97–100% of paired comparisons and DR loss in 95–96%; across the neural candidates, the corresponding ranges are 93–98% and 92–97%. Appendix Table 7 gives every prevalence and the stringent count of environments in which all 20 seeds reverse.

The fixed-candidate rows establish reversal without candidate switching. The selected EconML row answers a different question: R-loss continues to identify a competitive member of the pool while absolute pool accuracy deteriorates. Mean endpoint regret is .016, whereas mean oracle-best risk rises by about 1.31. Observable validation can therefore maintain good relative selection while rewarding a pool-wide movement away from the causal target.

5.3 How much can proxies help recover a hidden confounder?

The transfer study in Section 5.2 withholds U. The proxy experiment instead asks how validation behaves as pretreatment measurements contain progressively less information about it. CEVAE, CausalPFN, DRLearner, and causal forest receive three measurements $Z_k = U + \epsilon_k$. Three noise levels are compared with measurements independent of U, while the underlying (X, U, T, Y) draws remain paired. Estimators supplied with true U provide information references.

The central pattern is graded: informative proxies reduce causal deterioration, whereas weakening them turns monitored R-loss increasingly in the wrong direction. For fixed CEVAE, the severity-zero to-one root-PEHE increase progresses from .106 at proxy noise $\sigma = .5$, to .313 at $\sigma = 1$, .609 at $\sigma = 2$, and .900 with uninformative measurements. Over the same progression, monitoring R-loss changes by $+.005$, $-.014$, $-.154$, and $-.454$. Thus the score is initially nearly flat, then improves more strongly precisely as the estimator loses causal accuracy. Because these are fixed-CEVAE comparisons, the reversal cannot be attributed to candidate switching.

The result extends beyond CEVAE. CausalPFN, causal forest, DRLearner, and the selected candidate move from modest deterioration with informative proxies to large risk increases accompanied by falling R-loss when proxy information is poor. For the selected candidate, the root-PEHE change grows from $+.100$ to $+1.112$, while its R-loss change moves from $+.023$ to $-.451$. CausalPFN's predictive intervals also become less informative: the correlation between mean interval width and oracle error falls from $-.185$ at $\sigma = .5$ to $-.826$ with uninformative measurements. Thus proxy-aware models benefit from informative measurements, yet their validation and uncertainty outputs can become more reassuring as those measurements cease to identify the latent cause. Appendix F reports all conditions, comparators, paired intervals, true-U references, and uncertainty results.

6 Discussion and limitations

Our results expose a distinction that is easy to miss in CATE evaluation. An observable criterion may rank the candidates in a pool well even while every candidate moves farther from the causal target. Across the estimator classes and structural environments studied here, stronger hidden confounding typically raises oracle CATE error while lowering both R-loss and DR loss. The population construction explains this direction: once conditional exchangeability fails, the losses favor an observed-data contrast that can diverge systematically from the CATE. Low selection regret therefore answers a relative question—whether observable selection found a candidate close to the oracle-best member of the pool—but does not establish that either candidate is causally accurate. The proxy results sharpen the same point: additional measurements can reduce the damage from a latent cause, while architectural sophistication and predictive uncertainty need not reveal when the remaining information no longer supports accurate causal recovery.

The practical response is a severity-path validation audit. Before comparing methods, the developer specifies a scientifically plausible identification failure and a graded path that strengthens it. At each severity, candidates are refitted, selection uses an observable criterion, and that criterion is recomputed on independent monitoring data. Three trajectories are then reported together: the selected estimator's causal error, the oracle-best causal error in the pool, and the monitoring score. Rising oracle-best error signals that the pool or identification strategy must change; a widening gap between selected and oracle-best error identifies selection failure; and an improving score alongside rising causal error identifies directional misalignment of the criterion. Fixed-estimator trajectories provide the corresponding diagnosis without candidate switching.

The experiments study binary treatment, squared-error validation criteria, finite estimator pools, and proxy measurements within structural models where oracle causal error is available. These limits point directly to the next questions. Theory should characterize when criterion–error reversal is absent or unavoidable and guide criteria with explicit alignment guarantees. Empirical work should construct domain-grounded and semi-synthetic severity paths and test whether the audit predicts failure in realistic applications. Extensions to multivalued and continuous treatments, interacting identification failures, policy-value criteria, partial-identification procedures, and proximal estimators (Miao et al., 2018) would broaden the framework beyond the setting studied here.

AI USE STATEMENT

We used generative AI tools to assist with writing code and with minor grammar and wording edits. The authors reviewed and verified the resulting code and take responsibility for the final text, claims, code, figures, and analyses.

REPRODUCIBILITY STATEMENT

Section 3 gives the population construction and proof, and Section 4 specifies the structural equations, losses, data splits, and evaluation protocol. Appendices A–F record sample sizes, severity grids, paired seeds, nuisance specifications, estimator settings, uncertainty procedures, all 24 environment definitions, neural training details, and proxy-study settings. The accompanying anonymized analysis-reconstruction package contains machine-readable protocols, frozen cell-level and environment-level analysis inputs, exact released-implementation and checkpoint identifiers, and scripts that reconstruct Figure 1 and every numerical table while verifying hashes and planned coverage.

REFERENCES

Ahmed M. Alaa and Mihaela van der Schaar. Validating causal inference models via influence functions. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 191–201, 2019.

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

Vahid Balazadeh Meresht, Hamidreza Kamkari, Valentin Thomas, Junwei Ma, Bingru Li, Jesse C. Cresswell, and Rahul G. Krishnan. CausalPFN: Amortized causal effect estimation via in-context learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 154945–154984, 2025.

Keith Battocchi, Eleanor Dillon, Maggie Hei, Greg Lewis, Paul Oka, Miruna Oprescu, and Vasilis Syrgkanis. Econml: A python package for ml-based heterogeneous treatment effects estimation. GitHub repository, 2019. URL `https://github.com/py-why/EconML`.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep universal probabilistic programming. *Journal of Machine Learning Research*, 20(28):1–6, 2019.

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

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B*, 82(1):39–67, 2020.

Alicia Curth and Mihaela van der Schaar. On inductive biases for heterogeneous treatment effect estimation. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15883–15894, 2021.

Alicia Curth and Mihaela van der Schaar. In search of insights, not magic bullets: Towards demystification of the model selection dilemma in heterogeneous treatment effect estimation. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6623–6642, 2023.

Alexander D'Amour, Peng Ding, Avi Feller, Lihua Lei, and Jasjeet Sekhon. Overlap in observational studies with high-dimensional covariates. *Journal of Econometrics*, 221(2):644–654, 2021.

Vincent Dorie, Jennifer Hill, Uri Shalit, Marc Scott, and Dan Cervone. Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *Statistical Science*, 34(1):43–68, 2019. doi: 10.1214/18-STS667.

Jennifer L. Hill. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics*, 20(1):217–240, 2011. doi: 10.1198/jcgs.2010.08162.

Andrew Jesson, Sören Mindermann, Uri Shalit, and Yarin Gal. Identifying causal-effect inference failure with uncertainty-aware models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11637–11649, 2020.

Andrew Jesson, Sören Mindermann, Yarin Gal, and Uri Shalit. Quantifying ignorance in individual-level causal-effect estimates under hidden confounding. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4829–4838, 2021.

Fredrik D. Johansson. IncomeSCM: From tabular data set to time-series simulator and causal estimation benchmark. In *Advances in Neural Information Processing Systems*, volume 37, pp. 136173–136192, 2024.

Edward H. Kennedy. Towards optimal doubly robust estimation of heterogeneous causal effects. *Electronic Journal of Statistics*, 17(2):3008–3049, 2023. doi: 10.1214/23-EJS2157.

Sören R. Künzel, 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*, 116(10):4156–4165, 2019.

Marc Lipsitch, Eric Tchetgen Tchetgen, and Ted Cohen. Negative controls: A tool for detecting confounding and bias in observational studies. *Epidemiology*, 21(3):383–388, 2010.

Christos Louizos, Uri Shalit, Joris M. Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6446–6456, 2017.

Divyat Mahajan, Ioannis Mitliagkas, Brady Neal, and Vasilis Syrgkanis. Empirical analysis of model selection for heterogeneous causal effect estimation. In *International Conference on Learning Representations*, 2024.

Wang Miao, Zhi Geng, and Eric J. Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

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

Harsh Parikh, Carlos Varjao, Louise Xu, and Eric Tchetgen Tchetgen. Validating causal inference methods. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 17346–17358, 2022.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, second edition, 2009.

Peter M. Robinson. Root-N-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

Paul R. Rosenbaum. *Observational Studies*. Springer, second edition, 2002.

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

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974.

Yuta Saito and Shota Yasui. Counterfactual cross-validation: Stable model selection procedure for causal inference models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8398–8407, 2020.

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

Uri Shalit, Fredrik D. Johansson, and David Sontag. Estimating individual treatment effect: Generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3076–3085, 2017.

Claudia Shi, David M. Blei, and Victor Veitch. Adapting neural networks for the estimation of treatment effects. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jerzy Splawa-Neyman. On the application of probability theory to agricultural experiments. Essay on Principles. Section 9. *Statistical Science*, 5(4):465–472, 1990. doi: 10.1214/ss/1177012031. Translated and edited by D. M. Dabrowska and T. P. Speed from the 1923 Polish original.

Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242, 2018. doi: 10.1080/01621459.2017.1319839.

A Reproducibility protocol

Table 3 records the sample sizes and replication counts for each experiment. A cell fixes the environment or family, severity, seed, and proxy condition. Independent random streams generate its data splits. Within each split, the seed is reused across severities so that covariates, latent variables, potential outcomes, outcome noise, and treatment-assignment uniforms stay paired.

Table 3: Dataset sizes per cell, planned cell counts, and replication counts. The selection/validation column serves model selection; monitoring is independently sampled and used only for scoring.

Experiment	Cells	Seeds	Train	Select./valid.	Monitor	Oracle test
EconML, base families	660	30	2,000	1,000	2,000	5,000
Neural, base families	660	30	2,000	1,000	2,000	5,000
EconML, 24 environments	5,280	20	1,500	750	1,500	3,000
Neural, 24 environments	5,280	20	2,000	1,000	1,500	3,000
Proxy conditions	1,320	30	2,000	1,000	2,000	5,000

Sample sizes are fixed within estimator family because the released neural training protocols use 2,000 training observations and a 1,000-observation selection set for representation learning and early stopping. The environment-bank conclusions compare paired severity changes within each family, rather than absolute error levels across the EconML and neural families; monitoring and oracle-test sizes remain common across the two transfer analyses.

The proxy experiment uses $s \in \{0, .1, .2, \ldots, 1\}$. The base hidden-confounding families and the environment bank use $s \in \{0, .025, .05, .075, .1, .15, .2, .25, .5, .75, 1\}$. Each base pool has $2 \times 11 \times 30 = 660$ cells. Each transfer experiment plans $24 \times 11 \times 20 = 5{,}280$ hidden-confounding cells. All endpoint pairs are complete; the FlexTENet trajectory contains 5,277 cells after three intermediate-severity failed fits. The proxy study contains $4 \times 11 \times 30 = 1{,}320$ cells and six fitted estimators per cell.

EconML estimators. We use EconML 0.17.0 (Battocchi et al., 2019). The random-forest regressions have 200 trees, a minimum leaf size of 10, and a feature fraction of .8; propensity models use logistic regression with a maximum of 2,000 iterations. DRLearner, CausalForestDML, and NonParamDML use five shuffled folds with the same seed-based construction. CausalForestDML has 200 trees and a minimum leaf size of 10; DRLearner uses a minimum propensity of .02. XLearner uses its cross-arm imputation procedure (Künzel et al., 2019). All four return CATE predictions through their effect interfaces. R-loss evaluates these contrasts using a nuisance pair shared by all candidates in the cell.

DR-loss nuisances. The DR pseudo-outcome uses a logistic propensity and separate 200-tree random-forest outcome regressions for treated and control observations, all fitted on the training split and predicted on the monitoring covariates. Estimated propensities are clipped to $[.025, .975]$. The same nuisance predictions score every candidate in a cell, and the monitoring outcomes enter neither fitting nor selection.

Pairing and uncertainty. For a quantity a, the endpoint change is $\widehat{\Delta}_a = J^{-1} \sum_{j=1}^{J} (a_{1,j} - a_{0,j})$. Seed bootstrap draws resample the J seed identities with replacement and retain each sampled seed's full trajectory. The reported 95% intervals are percentile intervals. Environment summaries first average within each environment and use 10,000 environment-bootstrap draws for the reported environment-clustered intervals. The correlation controls in Appendix C use 500 seed-bootstrap draws.

Base designs and environment bank. The two base families instantiate the theoretical mechanism directly. The 24-environment balanced fractional design and estimator settings form a single evaluation protocol. The supplementary package contains the environment definitions, settings, implementation versions, and reported summaries.

Reproducibility. Every reported endpoint comparison covers its planned pairs, and training, selection, monitoring, and oracle-test samples use independent random streams. Appendix E reports neural fitting details. CEVAE uses Pyro's (Bingham et al., 2019) implementation; CausalPFN uses its official package and released checkpoint. Software versions and source/checkpoint hashes accompany the supplementary package.

B BASE-FAMILY ENDPOINT RESULTS

Table 4: Severity-zero-to-one changes for the R-loss-selected estimators (30 paired seeds). Root-PEHE intervals are paired by seed; R-loss is evaluated on the independent monitoring split.

Pool	Family	$\mathcal{R}_0$	$\mathcal{R}_1$	$\Delta\mathcal{R}$ (95% CI)	Δ R-loss
EconML	Additive	.374	1.411	1.036 (1.007,1.062)	$-.458$
EconML	Nonlinear	.368	1.422	1.054 (1.027,1.081)	$-.473$
Neural selection	Additive	.439	1.419	.979 (.943,1.014)	$-.469$
Neural selection	Nonlinear	.447	1.455	1.008 (.980,1.041)	$-.478$

Endpoint regret is .006 and .013 for the EconML additive and nonlinear families and .001 and .022 for the corresponding neural families. The oracle-best root-PEHE increases by 1.047, 1.048, .983, and 1.002, respectively. Paired 95% intervals for the neural R-loss changes are $[-.489, -.444]$ and $[-.499, -.459]$.

C R-LOSS NUISANCE AND SCALE CONTROLS

These controls hold the fitted effect estimators and their selection decisions fixed, then vary how R-loss is evaluated. Selection uses the practical nuisance specification throughout. On a monitoring sample of size n, the empirical score is

$$\widehat{L}_s(g) = \frac{1}{n}\sum_{i=1}^{n}\big[Y_i - \widehat{m}_s(X_i) - \{T_i - \widehat{e}_s(X_i)\}g(X_i)\big]^2. \tag{8}$$

Nuisance specifications. The practical specification fits a 200-tree random forest with minimum leaf size 10 for m and logistic regression for e on the training split. The cross-fitted specification uses five monitoring folds: gradient boosting estimates m and logistic regression estimates e, with each observation scored by models fitted on the other four folds. The oracle specification uses the simulator's observed-X conditional means. For hidden confounding, 64-node Gauss–Hermite quadrature integrates out U to obtain $m(x) = \mathbb{E}[Y \mid X = x]$ and $e(x) = \Pr(T = 1 \mid X = x)$. Thus the oracle score retains the same observed information set as the practical score.

Reference scores. At the same severity, seed, monitoring sample, and nuisance specification, we evaluate the fitted CATE, the true CATE τ, and the zero-effect function 0. The two scale controls are

$$E_s(g) = \widehat{L}_s(g) - \widehat{L}_s(\tau), \qquad C_s(g) = \frac{\widehat{L}_s(g) - \widehat{L}_s(\tau)}{\widehat{L}_s(0) - \widehat{L}_s(\tau)}. \tag{9}$$

E_s removes the loss level achieved by the true effect at that severity; C_s additionally expresses the excess relative to the zero-effect comparison. Ratios are evaluated when the denominator magnitude is at least 10^{-6}, with its sign retained. The simulator supplies τ for these explanatory controls. Under confounding, the population minimizer g_R can achieve lower R-loss than τ, as Equations 3 and 6 explain.

For the unnormalized score, we use the paired change $\Delta\widehat{L}_s(g) = \widehat{L}_s(g) - \widehat{L}_0(g)$ within each seed; this removes the seed-specific baseline level and isolates movement along the severity path. Table 5 reports Spearman correlations between this paired change (or each same-severity scale control) and oracle root-PEHE across the 300 nonzero-severity cells in each family. Every interval lies below zero: the inverse score–error relationship persists across nuisance specifications and both scale controls.

Table 5: Continuous score–error correlations under hidden confounding. Parentheses contain 95% intervals from resampling the 30 seeds. Each cell retains its original selected estimator.

Nuisances	Score	Additive	Nonlinear
Practical	Paired ΔR-loss	$-.917$ ($-.934$, $-.902$)	$-.863$ ($-.905$, $-.822$)
Cross-fitted	Paired ΔR-loss	$-.891$ ($-.910$, $-.868$)	$-.860$ ($-.897$, $-.827$)
Oracle	Paired ΔR-loss	$-.935$ ($-.949$, $-.921$)	$-.918$ ($-.944$, $-.892$)
Practical	Excess E_s	$-.920$ ($-.943$, $-.895$)	$-.912$ ($-.931$, $-.891$)
Cross-fitted	Excess E_s	$-.917$ ($-.937$, $-.898$)	$-.914$ ($-.937$, $-.889$)
Oracle	Excess E_s	$-.921$ ($-.939$, $-.900$)	$-.912$ ($-.931$, $-.895$)
Practical	Normalized C_s	$-.924$ ($-.946$, $-.901$)	$-.915$ ($-.937$, $-.895$)
Cross-fitted	Normalized C_s	$-.920$ ($-.939$, $-.901$)	$-.918$ ($-.942$, $-.893$)
Oracle	Normalized C_s	$-.924$ ($-.942$, $-.905$)	$-.918$ ($-.937$, $-.901$)

D Environment bank and transfer results

The bank varies covariate distribution, response surface, CATE, latent structure, and treatment assignment. Table 6 lists the 24 combinations. For environment k, potential outcomes have conditional means $b_k(X) + h_k(U) + t\tau_k(X)$, with the specified outcome-noise distribution. Writing $q_k(X)$ for its standardized treatment score and ℓ_k for its latent loading vector, assignment follows

$$\pi_{k,s}(X, U) = \text{logit}^{-1}\{a_k q_k(X) + d_k s\, \ell_k^\top U\}. \tag{10}$$

The design fixes a_k, d_k, loadings, coefficients, covariances, and nonlinearities. Assignment probabilities are clipped to $[10^{-4}, 1 - 10^{-4}]$. The severity path couples U increasingly strongly to treatment as s increases.

Table 6: Structural combinations in the environment bank. Covariates: CG correlated Gaussian, HT heavy-tailed Student-t, GM Gaussian mixture, NM nonlinear manifold, SK skewed, MB mixed binary/continuous. Response: SA sparse additive, SN smooth nonlinear, IR interaction-rich, TP threshold/piecewise. CATE: C constant, SL sparse linear, SN smooth nonlinear, SD subgroup. Latent: dimension 1/3 with additive/nonlinear link A/N. Treatment: L linear, N nonlinear, IR interaction-rich.

ID	Covariates	Response	CATE	Latent	Treatment
0	NM	SA	SL	1N	L
1	NM	IR	C	3N	L
2	GM	SN	SD	1A	N
3	GM	IR	SN	3N	IR
4	HT	TP	C	1N	IR
5	GM	SA	SD	3A	N
6	SK	TP	SN	3A	L
7	SK	SA	C	3A	L
8	HT	TP	SL	1A	IR
9	CG	IR	SL	1A	IR
10	CG	TP	SD	1A	IR
11	MB	SN	C	3A	L
12	CG	IR	SD	1A	IR
13	SK	SN	SN	3N	N
14	MB	SA	C	3A	N
15	CG	TP	SL	1N	IR
16	MB	IR	SL	3N	L
17	HT	SN	SN	1A	N
18	HT	SN	SD	3N	N
19	NM	SA	C	3A	L
20	GM	SA	SN	1N	L
21	SK	IR	SL	1N	N
22	MB	SN	SN	1N	IR
23	NM	TP	SD	3N	N

Endpoint comparison. A strict reversal has Δroot-PEHE> 0 and a negative score change. Because DR pseudo-outcomes can be heavy-tailed, Table 7 reports individual paired-cell prevalence and the

more stringent count requiring reversal for all 20 seeds. The effect-size summaries in Table 2 first average endpoint changes within each environment and then summarize their distribution across environments.

Table 7: Reversal prevalence across the 24 hidden-confounding environments. Each percentage is the fraction of 480 paired severity-zero-to-one comparisons with rising root-PEHE and falling monitored loss. Parentheses count environments in which all 20 endpoint pairs reverse.

Family	Candidate	R-loss	DR loss
EconML	XLearner	99.0% (22/24)	95.8% (18/24)
	DRLearner	97.1% (21/24)	94.6% (18/24)
	CausalForestDML	99.6% (22/24)	96.25% (17/24)
	NonParamDML	97.9% (22/24)	95.2% (18/24)
	Selected (four)	99.0% (22/24)	95.8% (17/24)
Neural	TARNet	96.9% (21/24)	94.8% (18/24)
	CFRNet	93.1% (19/24)	92.1% (18/24)
	DragonNet	97.9% (22/24)	96.9% (18/24)
	FlexTENet	97.3% (22/24)	96.7% (18/24)

Trajectory comparison. Across nonzero-severity cells, the selected EconML estimator's monitoring R-loss and oracle error have negative Spearman correlations in 23 environments, with all 23 seed-bootstrap intervals below zero. Environment 18 has a positive point correlation with an interval spanning zero. These trajectory correlations complement the endpoint comparison by using all intermediate severities.

E Neural fitting and independent monitoring

TARNet and CFRNet use the authors' released TensorFlow 0.12 implementation; DragonNet uses its authors' code with TensorFlow 1.15 and Keras 2.2.4. TARNet/CFRNet train for 3,000 Adam iterations with learning rate .001, batch size 100, three representation layers of width 200, and three output layers of width 100. The imbalance penalty is zero for TARNet and .3 for CFRNet, with Wasserstein balancing. DragonNet uses targeted regularization, an Adam stage of up to 100 epochs, and an SGD stage of up to 300 epochs. Early stopping uses a 20% internal holdout of the training sample; the external selection and monitoring samples retain their evaluation roles. FlexTENet uses the authors' JAX implementation with its published default configuration, 10,000 optimization iterations, early stopping, and a seed fixed by the experimental cell (Curth & van der Schaar, 2021). All four networks are evaluated by the same R-loss, DR-loss, and oracle root-PEHE definitions.

The base-family neural selection experiment contains 660 hidden-confounding cells and selects among TARNet, CFRNet, and DragonNet. Appendix B reports the selected candidate's monitored R-loss and oracle-test risk changes. The 24-environment audit evaluates all four networks as fixed candidates under both monitoring losses.

F Proxy information and predictive uncertainty

The proxy experiment uses the additive latent outcome contribution $h(U) = U$ from Section 4.1. Three independent measurement noises produce $Z_k = U + \epsilon_k$, $\epsilon_k \sim N(0, \sigma^2)$, for $\sigma = .5, 1, 2$. The uninformative condition instead supplies three independent standard-normal measurements. A separate random stream generates proxy noise, preserving the same (X, U, T, Y) at a fixed severity and seed across proxy conditions.

CEVAE uses Pyro 1.9.1 with a normal outcome distribution, latent dimension 5, two hidden layers of width 64, 100 epochs, batch size 100, and learning rate .001. Predictions average 100 posterior samples. CausalPFN uses package version 0.1.4 and its released `causalpfn_v0.pt` checkpoint. Its 4,096-observation context accommodates all 2,000 training observations. Point estimates use the official CATE interface; 95% intervals use 1,000 samples. Both methods receive $[X, Z]$, as do the DRLearner and causal-forest comparators.

Selection uses validation R-loss over the four observed-input candidates. A new 2,000-observation monitoring draw, independently seeded from training, validation, and oracle test, is used only to recompute R-loss after selection. Table 8 reports fixed-CEVAE endpoint changes on this monitoring sample; each row follows the same estimator across severity.

Table 8: Fixed-CEVAE severity-zero-to-one changes with 95% paired-seed intervals (30 seeds). Negative ΔR-loss means an improving independently monitored validation score.

Proxy condition	Δ root-PEHE	Δ monitoring R-loss
$\sigma = .5$	$+.106$ $[.074, .138]$	$+.005$ $[-.004, .014]$
$\sigma = 1$	$+.313$ $[.283, .344]$	$-.014$ $[-.026, -.002]$
$\sigma = 2$	$+.609$ $[.569, .648]$	$-.154$ $[-.167, -.141]$
Uninformative	$+.900$ $[.847, .955]$	$-.454$ $[-.471, -.437]$

The three observed-input comparators also deteriorate in every condition. Table 9 reports their paired endpoint changes together with the validation-selected candidate.

Table 9: Proxy-study severity-zero-to-one changes. Each entry is Δroot-PEHE / Δ monitoring R-loss; selection uses validation R-loss over the four observed-input candidates.

Estimator	$\sigma = .5$	$\sigma = 1$	$\sigma = 2$	Uninformative
CEVAE	$.106/.005$	$.313/-.014$	$.609/-.154$	$.900/-.454$
CausalPFN	$.078/.021$	$.355/-.000$	$.776/-.146$	$1.125/-.455$
Causal forest	$.090/.018$	$.300/-.001$	$.669/-.144$	$1.025/-.453$
DRLearner	$.439/.045$	$.449/.011$	$.622/-.145$	$.922/-.466$
Selected	$.100/.023$	$.336/-.002$	$.758/-.143$	$1.112/-.451$

DRLearner and a regression estimator supplied with true U provide information references. Their root-PEHE changes from $.438$ to $.823$ and from $.461$ to $.838$. Increasing severity also concentrates treatment assignment conditional on (X, U), so these reference trajectories record the finite-sample difficulty of estimating effects under the same assignment path even when U is observed. They are shared across proxy conditions because their inputs are $[X, U]$.

Interval width. For each fixed CausalPFN fit, we average the pointwise 95% interval widths and compare this quantity with oracle root-PEHE across severity and seed. Table 10 shows negative correlations in every proxy condition, becoming stronger as measurements lose information. The intervals describe uncertainty under CausalPFN's pretrained model; the stress test measures how that output responds when the identifying information changes.

Table 10: Spearman correlation between CausalPFN's mean interval width and oracle root-PEHE, with 95% seed-bootstrap intervals over all 11 severities and 30 seeds.

Proxy condition	Correlation	95% interval
$\sigma = .5$	$-.185$	$[-.341, -.051]$
$\sigma = 1$	$-.463$	$[-.613, -.350]$
$\sigma = 2$	$-.686$	$[-.792, -.588]$
Uninformative	$-.826$	$[-.890, -.766]$