CAUSAL MATCHING WITH A RANK-LEARNED PROGNOSTIC SCORE

Anonymous authors
Paper under double-blind review

ABSTRACT

Can pairwise outcome order suffice to learn useful causal matches? We answer this question by fitting a scalar prognostic score from binary comparisons of observed historical control outcomes, then using it for matching in an independent future cohort. This learner discards outcome magnitudes and is invariant to strictly increasing transformations of historical outcomes, with the fitting procedure fixed. Under a Gumbel location model, comparison probabilities calibrate prognosis differences up to scale. Our matching-specific bound transfers score error to mean squared error (MSE) of the average treatment effect on the treated (ATT), accounting for future support, adaptive donor selection and reuse. Standard ranking theory provides individual-level learning guarantees under explicit finite-class conditions. Six outcome-noise settings and an independently seeded sample-size study delimit the empirical benefit. In simulation study, increasing historical fitting size from 300 to 2400 changes RankNet from nearly tied with the state of art method to up to 90.0% lower mean conditional design MSE. The results establish effective matching from coarsened outcome supervision in this setting, while also showing the advantage of flexible full-outcome regression.

1 INTRODUCTION

Matching estimates treatment effects by using untreated patients as counterfactual donors. Covariate proximity need not imply similar untreated prognosis: distant covariate values can lie on the same level set of a nonlinear response surface. A learned diagonal distance can miss this structure, whereas a flexible scalar prognostic score can join patients with similar expected outcomes.

We retain pairwise order from a completed historical cohort: $B_{ij} = \mathbf{1}\{Y_i > Y_j\}$ among controls. A scalar RankNet score learned from these comparisons is frozen before matching an independent future cohort. Full-outcome regressors use the same patients but also retain outcome magnitudes. Comparisons coarsen existing outcomes; they add neither patients nor expert knowledge. Separate fitting prevents future outcomes from selecting matches, extending pilot-sample prognosis fitting (Aikens et al., 2020).

This distinction gives the procedure a concrete invariance: strictly increasing transformations of historical fitting and validation outcomes leave the pair labels unchanged. With fixed pair draws, initialization, optimization and rank-based validation, they leave the fitted score and future donor assignment unchanged. Numerical-outcome regression need not have this property. This invariance concerns historical supervision; future outcomes and the ATT scale remain fixed.

When does this coarsened supervision suffice? Order alone does not identify cardinal prognosis gaps. We give location-noise conditions for logistic calibration and a matching-specific ATT mean squared error (MSE) bound separating future support, score error and adaptive donor reuse. The experiments assess what is gained and lost by retaining only comparisons.

In an independently seeded study of folded prognosis geometry, increasing historical sample size reveals an advantage over MALTS. We also evaluate full-outcome historical learners, future-outcome BART and causal forest, and a semi-synthetic LaLonde scale comparison.

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.

1.1 Related work

Prognostic and propensity scores summarize outcome and treatment information, respectively (Rosenbaum & Rubin, 1983; Hansen, 2008; Stuart, 2010); joint-score matching includes empirical and doubly robust constructions (Leacy & Stuart, 2014; Antonelli et al., 2018). MALTS learns outcome-adaptive distances, Variable Importance Matching permits modular outcome learners, and Neural Score Matching uses neural propensity representations (Parikh et al., 2022; Lanners et al., 2023; Clivio et al., 2022). PairNet learns from observed outcome contrasts (Nagalapatti et al., 2024). Our contribution concerns the specified comparison score and its donor-assignment MSE, not outcome-adaptive or pair-based causal learning in general.

The closest ordinal precedent is Li & Fu (2017): ordinalized outcomes guide a kernel representation through ordinal scatter and distribution balancing, followed by nearest-neighbor ATT matching. We instead analyze scalar logistic-score calibration and capped donor assignment. Counterfactual representation bounds link prediction and distribution discrepancy for individual effects (Shalit et al., 2017); our target is ATT MSE.

Bradley–Terry, RankNet and preference-based reward modeling use logistic score differences (Bradley & Terry, 1952; Burges et al., 2005; Christiano et al., 2017). RankSVM and survival ranking learn from pairwise order (Joachims, 2002; Raykar et al., 2007). Active ranking studies query savings under geometric or sequential-noise assumptions (Jamieson & Nowak, 2011; Heckel et al., 2019); our pair budget is computational. Regression with comparisons uses additional ordinal information (Xu et al., 2020); our comparisons coarsen existing outcomes.

2 Outcome comparisons and future matching

2.1 Cohorts, target and assumptions

Let $X \in \mathbb{R}^p$ be pretreatment covariates, $A \in \{0,1\}$ treatment, and $Y(a)$ potential outcomes. An independent historical sample provides fitting and patient-level validation data. Historical controls used for comparison fitting have covariate law P. In the future cohort, let $\mathcal{T}$ and $\mathcal{C}$ index treated and control units, with sizes $N_1, N_0 > 0$. Under consistency and future conditional mean exchangeability, write $m_a(x) = \mathbb{E}[Y(a) \mid X = x]$. Our target is the sample conditional ATT,

$$\tau_X = \frac{1}{N_1} \sum_{i \in \mathcal{T}} \{m_1(X_i) - m_0(X_i)\}. \tag{1}$$

With constant treatment effects, this target equals population ATT; the LaLonde experiment below allows heterogeneous effects.

We use the following conditions explicitly. **A1 (consistency):** $Y = Y(A)$ and no interference between units. **A2 (future mean exchangeability):** $\mathbb{E}[Y(a) \mid X, A] = m_a(X)$ for $a = 0, 1$. **A3 (support and feasibility):** future arm-specific covariate laws have density ratios relative to historical control law P bounded by ω_1, ω_0, and $rN_0 \geq N_1$. **A4 (transport):** the historical and future untreated conditional mean is the same m_0. **A5 (comparison calibration):** historical control errors are iid centered Gumbel with constant scale s, independent of X. A1–A4 concern causal interpretation and transport; A5 is a sufficient statistical model for cardinal score calibration, not an identification assumption for ATT by itself.

2.2 Learners and information access

The families in Table 1 differ in what they retain from history. Mahalanobis uses covariates; propensity matching also uses treatment labels. Ridge, neural regression and boosting fit numerical control outcomes. Huber-MLP uses the same historical controls and hidden-layer width as RankNet, but retains their full outcome magnitudes. MALTS fits p diagonal stretches by minimizing within-arm neighbor outcome reconstruction plus a squared-stretch penalty, using numerical outcomes in both arms. RankNet uses only binary control comparisons. Oracle labels and numeric oracle prognosis are unavailable benchmarks. BART and causal forest, introduced in Section 4, additionally fit future-cohort outcomes and do not use donor assignment.

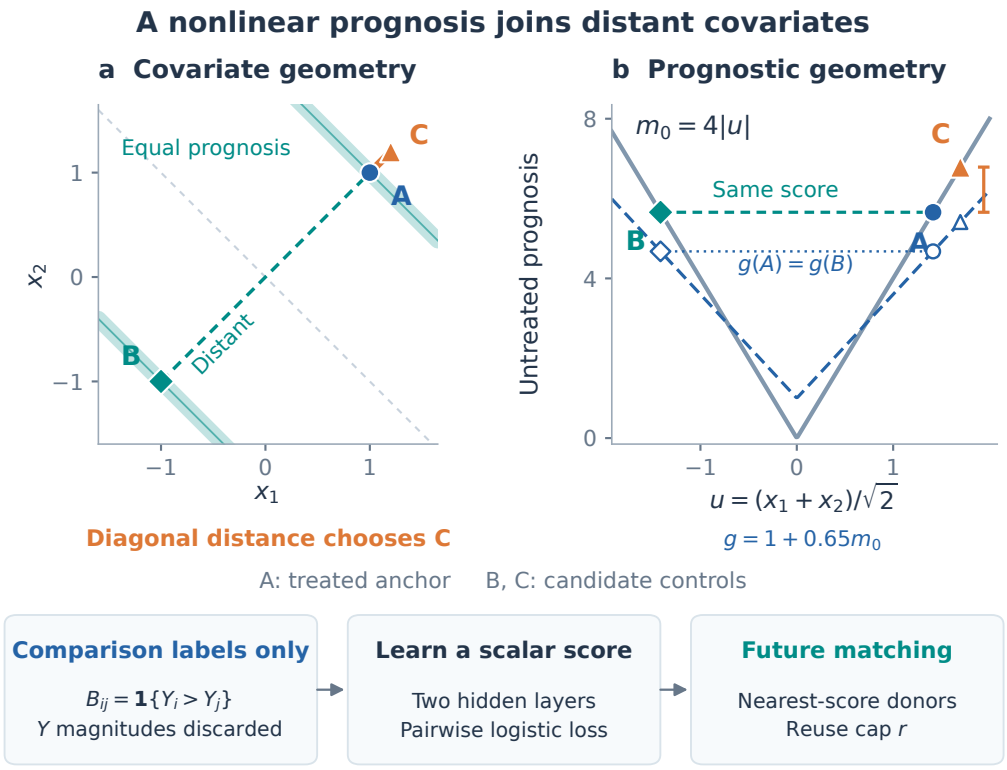

Figure 1: Why prognosis geometry can differ from covariate proximity. In this exact illustration, $m_0(x) = 4|(x_1 + x_2)/\sqrt{2}|$. Patients A and B have equal prognosis despite being farther apart than A and C under every diagonal metric with positive total weight. The dashed score is the illustrative affine calibration $g = 1 + 0.65m_0$, not a fitted result: it preserves equal prognosis for A and B. The workflow separates historical comparisons from future matching.

2.3 Comparison learner

Standardize covariates using historical fitting controls: $z(x)_k = (x_k - \bar{x}_k)/s_k$. For independent historical controls, define $B_{ij} = \mathbf{1}\{Y_i > Y_j\}$. We fit a scalar neural score $g_\theta(z)$ by minimizing

$$\widetilde{U}_M(\theta) = \frac{1}{M} \sum_{(i,j)\in\mathcal{P}_M} \left[\log\{1 + \exp(d_{ij})\} - B_{ij} d_{ij}\right], \qquad d_{ij} = g_\theta(z_i) - g_\theta(z_j). \tag{2}$$

Here $\mathcal{P}_M$ consists of sampled nonself pairs. The model is RankNet: its comparison probability is $\mathrm{sigmoid}(d_{ij})$. In the implementation, $g_\theta(z) = 8\tanh\{v^\top \mathrm{ReLU}(W_2\,\mathrm{ReLU}(W_1 z + b_1) + b_2)/8 + b/8\}$, with two hidden layers of width 32. The score is scalar and bounded, so it cannot exactly represent an unbounded calibrated target; θ comprises the displayed weights and biases.

2.4 Future donor assignment

For a prespecified reuse cap r with $rN_0 \geq N_1$, let $\mathcal{J}_r$ be the donor maps $J : \mathcal{T} \to \mathcal{C}$ satisfying $K_j(J) = \sum_{i\in\mathcal{T}} \mathbf{1}\{J(i) = j\} \leq r$. Select

$$\widehat{J}_r \in \arg\min_{J\in\mathcal{J}_r} \frac{1}{N_1} \sum_{i\in\mathcal{T}} \{\widehat{g}(X_i) - \widehat{g}(X_{J(i)})\}^2, \qquad \widehat{\tau} = \frac{1}{N_1} \sum_{i\in\mathcal{T}} (Y_i - Y_{\widehat{J}_r(i)}). \tag{3}$$

We absorb the fixed standardization into $\widehat{g}(x)$. Replicating each control r times reduces the assignment to a rectangular linear-sum assignment problem. The tie rule is deterministic and uses no future outcomes. For $r = N_1$, this is ordinary nearest-score matching with replacement. A smaller cap limits variance from donor reuse but can force larger prognosis gaps.

Table 1: Learners and information access. z is the covariate vector standardized by historical control fitting means and standard deviations. Every displayed score or distance uses the same assignment rule. MALTS uses raw coordinates in the synthetic studies and standardized coordinates in LaLonde.

Method	Score or squared distance	Learner and supervision
Mahalanobis	$(z-z')^\top \widehat{\Sigma}^{-1}(z-z')$	Covariance of all fitting covariates, pseudoinverse.
Propensity	$[\widehat{e}(z)-\widehat{e}(z')]^2$	Logistic regression, $C=1$, all fitting treatment labels.
Ridge	$[\widehat{m}(z)-\widehat{m}(z')]^2$	Linear regression, penalty 1, fitting control outcomes.
MLP squared/Huber	$[\widehat{m}(z)-\widehat{m}(z')]^2$	Neural network with two hidden layers of width 32; squared or Huber loss using numerical control outcomes, including magnitudes.
Boost squared/absolute	$[\widehat{m}(z)-\widehat{m}(z')]^2$	Histogram gradient boosting; squared or absolute outcome loss on controls.
MALTS	$\sum_{k=1}^{p} M_k^2(x_k-x_k')^2$	Original within-arm neighbor outcome reconstruction plus $\sum_k M_k^2$; p stretches, numerical outcomes in both treatment arms.
RankNet	$[g_\theta(z)-g_\theta(z')]^2$	Bounded scalar neural network, two hidden layers of width 32; binary control outcome comparisons only.
Oracle labels	$[g_\theta(z)-g_\theta(z')]^2$	Same RankNet trained on true m_0 comparisons; unavailable.
Oracle prognosis	$[m_0(x)-m_0(x')]^2$	True numerical prognosis; unavailable.

3 From comparison error to matching MSE

We separate identification of a score from its estimation and from donor selection. This matters because correct order does not generally identify cardinal closeness: both prognosis triples $(0,1,3)$ and $(0,2,3)$ have the same order, but the closest endpoint to the middle unit changes.

3.1 Calibration and its boundary

Proposition 1 (Calibration under location noise)**.** *Suppose historical controls satisfy $Y = m_0(X) + s(G-\gamma)$, where $s>0$, G is standard Gumbel independent of X, and $\gamma = \mathbb{E}G$. Then*

$$\mathbb{P}(Y > Y' \mid x, x') = \mathrm{sigmoid}\{[m_0(x)-m_0(x')]/s\}.$$

Over a class containing $g_* = m_0/s$, every population logistic-risk minimizer with finite second moment equals $g_* + c$ almost surely. Its matching distances therefore recover prognosis distances up to scale.

This is the standard Bradley–Terry random-utility calibration, specialized to observed control outcomes. With iid additive noise independent of covariates and having finite mean zero, majority order still follows m_0 when the difference-noise CDF is continuous and strictly increasing; the fitted logistic score need not be affine in m_0. Covariate-dependent skew can reverse mean order entirely. For example, $Y_a = 1$ with probability 0.9 and -10 otherwise has mean -0.1, yet exceeds $Y_b \equiv 0$ with probability 0.9. These are substantive restrictions on mean-prognosis matching, not just technical qualifications.

3.2 Learning a calibrated score

Define the centered error $e_g = g - g_* - \mathbb{E}_P(g-g_*)$. The transfer result takes as input a historical fit satisfying

$$\mathbb{E}_P e_{\widehat{g}}(X)^2 \le \varepsilon^2.$$

Ranking generalization is governed by independent patients, even when their outcomes generate many pairs (Clémençon et al., 2008; Agarwal & Niyogi, 2009). Pair subsampling adds a separate

computational approximation (Clémençon et al., 2016). Appendix A.2 gives a finite-class bound, its constants and a sampled-pair version with optimization error. Those conditions do not automatically hold for an early-stopped continuous neural network.

3.3 MATCHING-SPECIFIC MSE TRANSFER

The exact conditional MSE separates squared average prognosis mismatch from treated-outcome and reused-donor variance (Appendix A.3). We connect score error to this MSE while retaining adaptive donor selection.

Theorem 1 (Learning-to-matching transfer with a donor cap)**.** *Condition on positive future arm sizes, assume independent future units, and suppose their treated and control covariate laws have density ratios relative to P bounded by ω_1, ω_0. Assume stable untreated prognosis $m_0 = sg_* + c$, future mean exchangeability and conditional outcome variances at most v. Define*

$$A_r = \mathbb{E}_F \min_{J \in \mathcal{J}_r} \frac{1}{N_1} \sum_{i \in \mathcal{T}} \{m_0(X_i) - m_0(X_{J(i)})\}^2, \qquad C_r = \omega_1 + \frac{rN_0}{N_1}\omega_0.$$

For any historical fit with $\mathbb{E}_P e_{\widehat{g}}^2 \le \varepsilon^2$, the assignment in (3) *satisfies*

$$\mathbb{E}_F[(\widehat{\tau} - \tau_X)^2 \mid H] \le \left(\sqrt{A_r} + 2s\sqrt{2C_r\varepsilon^2}\right)^2 + \frac{v(1+r)}{N_1}. \tag{4}$$

Under the finite-class assumptions in Appendix A.2, take $\varepsilon^2 = a_n/\kappa$ with historical probability at least $1 - \delta$.

Corollary 1 (Consistency with vanishing matching and learning error)**.** *Consider a sequence satisfying Theorem 1, with feasible caps r_n. Suppose $\mathbb{E}_P e_{\widehat{g}_n}^2 \le \varepsilon_n^2$ on historical events of probability at least $1 - \delta_n$, where $\delta_n \to 0$, and*

$$A_{r_n} \to 0, \qquad s_n^2 C_{r_n} \varepsilon_n^2 \to 0, \qquad \frac{v_n(1+r_n)}{N_{1,n}} \to 0.$$

Then the conditional future MSE converges to zero in probability over historical samples, and $\widehat{\tau}_n - \tau_{X,n} \to 0$ in probability over both cohorts. Unconditional MSE convergence additionally requires control of the complementary historical events, for example uniform integrability of the conditional MSE.

Conditional Markov's inequality applied to the vanishing bound proves the corollary. With fixed cap and bounded $s, v, \omega_0, \omega_1, N_0/N_1$, it suffices that $N_1 \to \infty$, $\varepsilon_n \to 0$ and $A_r \to 0$. The last condition is substantive: support alone need not provide sufficiently similar donors under a fixed reuse constraint.

The first term separates available prognosis overlap from learned-score error. The second displays the price of reusing noisy controls. The key selection step is that every capped donor map, including one chosen from all future covariates, obeys

$$\frac{1}{N_1}\sum_i (e(X_i) - e(X_{J(i)}))^2 \le \frac{2}{N_1}\sum_{i \in \mathcal{T}} e(X_i)^2 + \frac{2r}{N_1}\sum_{j \in \mathcal{C}} e(X_j)^2. \tag{5}$$

Thus a selected donor need not be treated as an independent draw. This is the matching-specific connection; neither the underlying ranking rate nor the MSE decomposition is claimed as new. Increasing the cap enlarges the feasible donor set and can decrease A_r, while increasing both the variance bound and the score-error multiplier C_r. Poor future support makes the density-ratio term large or invalid. Without a cap, $r = N_1$ gives a weak bound, exposing the limitation rather than removing it.

Proposition 2 (Sensitivity to extreme historical outcomes)**.** *For scores bounded by B, replace at most k outcomes among n historical units, keeping covariates fixed. With $q_{n,k} = 1 - (n-k)(n-k-1)/\{n(n-1)\}$, the complete pairwise risk changes by at most $2Bq_{n,k}$ uniformly over scores. Its new minimizer has original empirical risk at most $4Bq_{n,k}$ above the original minimum.*

Only pairs touching replaced outcomes can change labels, and a label flip changes logistic loss by at most $|g(x) - g(x')| \le 2B$. The bound is independent of replacement magnitude. Squared outcome loss has no analogous magnitude-independent bound on unbounded outcomes. This comparison motivates a stress test, not a guarantee of lower matching MSE; Huber and absolute-loss regression are essential controls.

Table 2: True prognosis and propensity. $\ell(x)$ is the propensity logit before clipping; d_m, d_e count covariates entering the true models.

Family	p	d_m	d_e	$m_0(x)$; $\ell(x)$
Linear	10	4	3	$x_1 + 2x_2 - x_3 + 0.5x_4$; $0.5x_1 - x_2 + 0.3x_3$
Folded	10	2	1	$4\lvert x_1\rvert + 0.5x_2$; $2x_1 + 2(\lvert x_1\rvert - 0.8)$
Nonlinear	15	7	4	$2\sin(x_1x_2)+(x_3-1)^2+3\mathbf{1}\{x_4x_5>0\}+x_6x_7$; $0.5x_1-x_2+0.5x_3x_4$

4 EXPERIMENTS

We evaluate six synthetic settings, historical sample size, and semi-synthetic LaLonde data. Matching learners are trained on history; BART and causal forest are separately marked future-outcome references.

4.1 SIX-SETTING SYNTHETIC COMPARISON

Each of three prognosis families and two historical noise conditions has twenty repetitions, with 600 fitting, 300 validation and 800 future patients. All thirteen methods use the same future patients and realized outcomes. The primary endpoint is mean realized squared ATT error; Monte Carlo SE (MCSE) is the repetition-level standard deviation divided by $\sqrt{20}$. Paired contrasts use per-repetition differences. This endpoint accommodates estimators fitted on future outcomes; conditional design MSE (10) remains a separate matching-only diagnostic (Appendix B). No inference or coverage is evaluated.

DGPs and mechanism. Let $X \sim N(0, I_p)$, $Y(1) = Y(0) + 1$, and $A \mid X \sim \text{Bernoulli}\{e(X)\}$. Table 2 gives the true models. Propensities are clipped to $[0.02, 0.98]$. Historical noise is centered Gumbel with scale 0.5. In the contaminated condition it also contains an independent $I(U < 0.05)Z$, $Z \sim N(0, 20^2)$. Future outcome noise is $N(0, 0.5^2)$ in both conditions. Consequently, contamination represents poorer historical outcome measurement, while mean prognosis is stable. The clean model satisfies Proposition 1; contamination preserves majority order but changes its logistic link. Unbounded covariates and neural optimization fall outside the bounded finite-class guarantee.

For the folded component, covariate-distant points on opposite branches share prognosis. A diagonal metric can prefer a nearer point with a larger prognosis gap, whereas a nonlinear scalar score can connect the branches (Figure 1). This is a representational possibility, not a universal failure of MALTS.

Comparators and fitting. Matching methods use cap 5, with 1NN as a sensitivity analysis. Neural methods share hidden layers, initialization and patient-level validation. BART (Chipman et al., 2010) fits future controls and averages treated outcomes minus predicted untreated outcomes. Causal forest (Athey et al., 2019) uses future outcomes and an out-of-bag augmented inverse-probability ATT estimate. These are broader-information references, not historically trained matchers; neither uses true prognosis. Appendix B.2 specifies both estimators.

Results. On the clean fold, RankNet MSE is 1.589 versus MALTS 2.103; their paired difference is -0.514 (MCSE 0.183). Contaminated linear history gives 0.0315 versus 0.2021 for squared-loss MLP, while Huber-MLP reaches 0.0177 using full numerical historical outcomes. Clean linear history favors Ridge (0.0026 versus RankNet 0.0066). The contaminated fold is adverse: RankNet 3.381 exceeds absolute-loss boosting 1.915 and MALTS 2.640. In the nonlinear family, RankNet is 0.0529/0.0700 for clean/contaminated history versus MALTS 0.0427/0.0410; paired differences are smaller than two MCSEs, which does not establish equivalence.

Future-outcome references improve performance on the folded and nonlinear settings: on the fold, BART reaches 0.323 (MCSE 0.075) and causal forest 0.013 (0.004), versus RankNet's 1.589/3.381. Their additional access to the future cohort's uncontaminated outcomes matters when historical measurements are degraded. Numeric oracle prognosis also remains stronger than learned ordinal

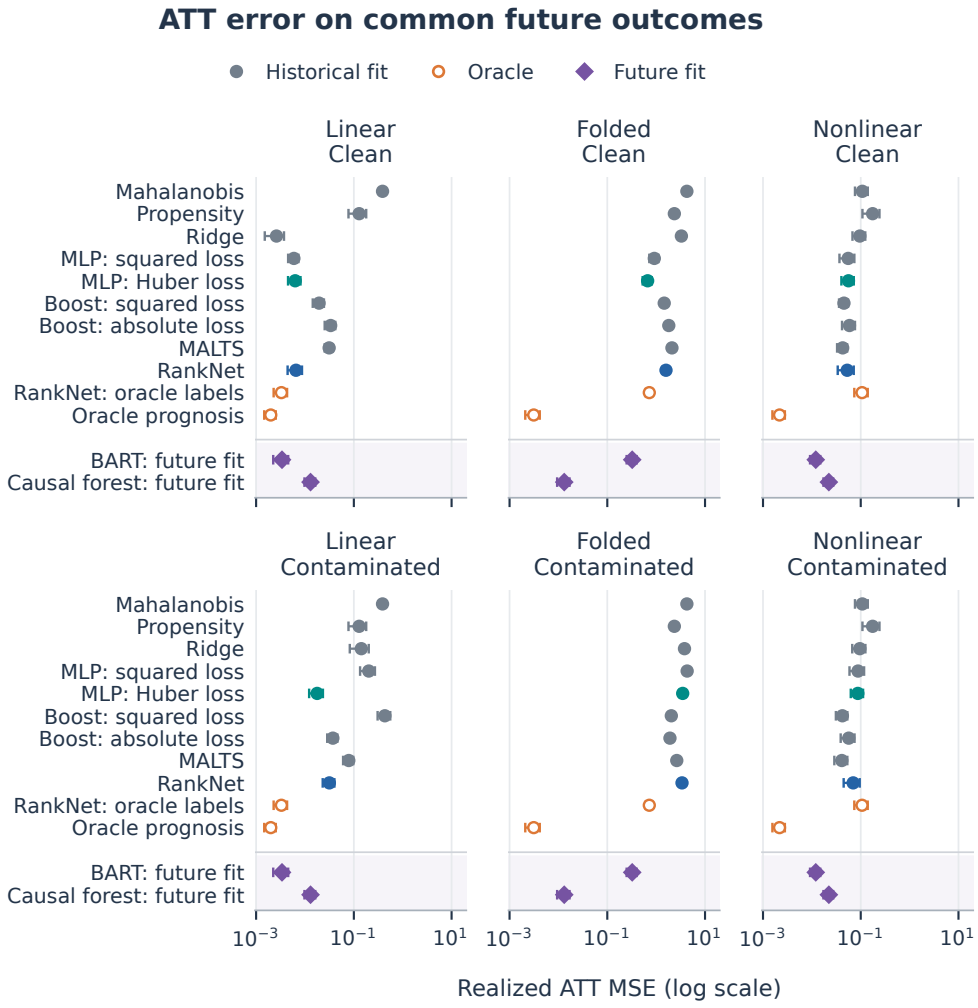

Figure 2: RankNet gains over MALTS in linear and clean folded settings, but loses on the contaminated fold; future-outcome references offer substantial further gains. All points are mean realized squared ATT error over twenty paired repetitions, with one MCSE. Historical matchers share cap 5; BART and causal forest additionally fit future outcomes. Oracle benchmarks are unavailable. Clean/contaminated histories share future data, so the future-outcome references repeat within each family. Axes are logarithmic.

supervision in nonlinear settings. The donor cap controls reuse variance in the theory; it is not uniformly best empirically.

4.2 Historical sample size and prognosis geometry

We chose a folded prognosis geometry in a separate pilot and fixed the following design before evaluation on disjoint confirmation seeds.

Independent confirmation. We fix $m_0(x) = 4|u_4| + 0.5x_{10}$ with $u_4 = \sum_{k=1}^{4} x_k/2$, $X \sim N(0, I_{10})$, and propensity logit $2u_4 + 2(|u_4| - 0.8)$, clipped as before. Five covariates enter prognosis and four enter propensity. Opposite projection branches share the folded component of prognosis, while treatment favors the positive branch. Twenty new seeds use nested historical fitting sizes $n \in \{300, 600, 1200, 2400\}$, validation size $n/2$, and the same 800 future patients within each seed. Width 32 is primary; widths 8 and 64 probe capacity. All neural models use up to 4000

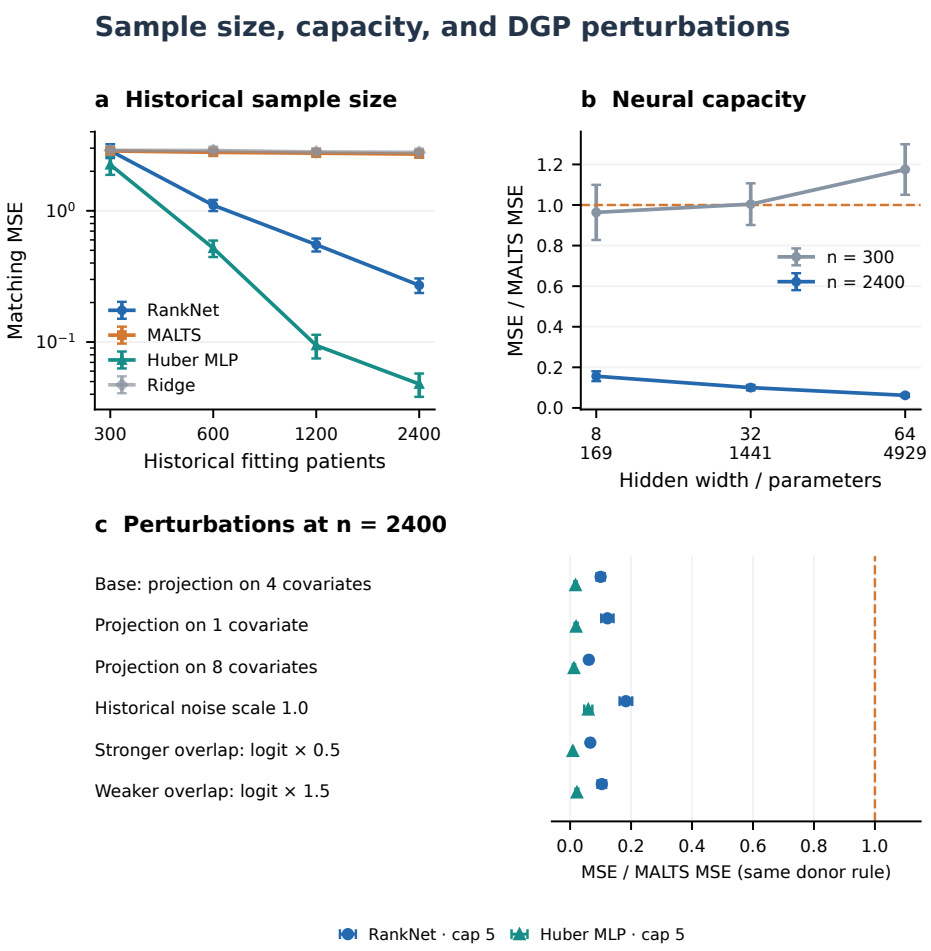

Figure 3: Historical sample size improves rank-learned matching relative to MALTS; full-outcome Huber regression remains stronger. All panels use cap 5 and twenty seeds; each seed shares 800 future patients across fitting sizes. Panels a and b show historical-size and network-width effects. Panel c varies projection dimension (1 or 8 covariates), historical noise (scale 1), or propensity-logit multiplier (0.5 or 1.5); it uses paired-seed jackknife errors for learner/MALTS ratios of means. Bars show one MCSE.

updates with validation stopping; MALTS uses three predetermined starts. Five one-factor perturbations at $n = 2400$ vary projection dimension, historical noise or propensity strength (Figure 3); Appendix B.1 gives the complete 180-design protocol.

Sample-size result (conditional design MSE). RankNet and MALTS are nearly tied at $n = 300$: mean MSE 2.869 (MCSE 0.347) and 2.857 (0.141), respectively. At $n = 2400$, RankNet reaches 0.271 (0.034), versus MALTS 2.706 (0.156), a 90.0% reduction; the paired difference is -2.435 (MCSE 0.155). Huber MLP, using full numerical historical control outcomes, is better at 0.048 (0.010), as is squared-loss MLP at 0.058 (0.010). The rank model therefore learns effective matches from order alone as historical information increases; full numerical outcomes permit still lower MSE.

Capacity and interpretation. At $n = 300$, width 64 has mean MSE 3.357 (MCSE 0.370), versus 2.753 (0.422) for width 8. At $n = 2400$, this ordering reverses: 0.168 (0.020) versus 0.423 (0.064). Width 32 remains the prespecified primary method.

Perturbations and donor-rule boundary. Under cap 5, the RankNet/MALTS mean-MSE ratio remains between 0.061 (MCSE 0.009; eight-coordinate projection) and 0.183 (0.021; doubled his-

torical noise) across all five perturbations (Figure 3). Huber MLP has lower mean MSE in every condition. Ordinary 1NN preserves the base oblique-fold advantage: RankNet MSE 0.273 versus MALTS 1.843, paired difference -1.570 (MCSE 0.100). However, the axis-aligned perturbation reverses the mean ordering: 0.219 versus 0.149, difference $+0.070$ (MCSE 0.055). Thus the cap can magnify a comparator's matching bias; the all-perturbation gain is specifically a cap 5 result. Doubled historical noise also changes the calibrated score range under the fixed bound; it does not isolate robustness.

4.3 Semi-synthetic LaLonde: heterogeneous effects and outcome scale

We use the 614-unit LaLonde sample distributed with MatchIt 4.7.2 (Dehejia & Wahba, 1999; Ho et al., 2011): eight covariates, 185 treated and 429 controls. A fixed control-earnings regression defines synthetic untreated means. Effects vary with standardized education; independent uniform noise gives nonnegative dollar outcomes. Twenty random splits allocate 246 patients to fitting, 61 to validation and 307 to future evaluation, including 93 future treated patients. Each estimate is evaluated against those patients' mean true effect. The experiment conditions on one empirical covariate pool; it does not estimate the original job-training effect.

Only historical supervision changes between dollars and $\log(1 + Y)$; all future outcomes and ATT targets remain in dollars. RankNet labels and hence fixed-seed fits and donors are unchanged. Figure 4 reports the complete comparison. In squared thousands of dollars, RankNet MSE is 0.393 (MCSE 0.087), Huber-MLP using dollar outcomes is 0.270 (0.111), and MALTS is 0.191 (0.064) on dollars and 0.120 (0.034) on log outcomes. BART and causal forest reach 0.797 (0.132) and 0.357 (0.070). Thus exact historical-scale invariance does not imply the lowest MSE. Appendix B.3 gives the generator and fitting protocol.

5 Discussion

A rank model makes the retained information explicit: fixed comparisons produce identical future matches under strictly increasing historical outcome transformations. This is useful when only order is retained or historical outcome scales differ. It requires no expert labels. The sample-size study shows an advantage over diagonal matching on the selected fold, while full-outcome Huber-MLP achieves lower MSE. BART and causal forest address the broader regime where future outcomes can train the estimator.

The limits are specific. A5 excludes noise patterns that break cardinal calibration or mean order. Bounded output limits approximation of unbounded prognosis; an output-bound ablation remains unrun. The selected geometries and one LaLonde generator do not establish broad superiority, and cap-specific gains can reverse under 1NN. Architecture and optimization differ, so parameter count alone is not a causal explanation. Finite-class theory does not certify fitted neural optimization; approximation and norm-based bounds offer related tools (Barron, 1993; Bartlett et al., 2017). Further comparisons should include other learned matchers and robust regression (Friedman, 2001), while varying output bounds, future donor availability and calibration beyond A5. We study MSE, not treatment-effect inference.

Reproducibility statement

Each evaluation uses twenty paired seeds; sample-size pilot and confirmation seeds are disjoint. Saved patients, fits, donors and hashed sources identify runs. Separate computational audits reconstruct MSE and assignment optimality. Appendix proofs and implementation details state the conditions. LaLonde uses public covariates and generated outcomes; no clinical data are used.

AI-USE STATEMENT

We used LLM-based AI tools to identify and summarize related literature, develop research hypotheses and simulation designs, implement methods and experiments, assist with mathematical derivations and proofs, interpret results, and draft and revise manuscript text and figures. AI assistance also included designing data-generating processes and writing code that generated synthetic datasets programmatically. AI-assisted computational checks and separate-agent reviews supported verification; these do not constitute independent human review. The authors take responsibility for the final manuscript, claims, code, and results.

REFERENCES

Alberto Abadie and Guido W. Imbens. Large sample properties of matching estimators for average treatment effects. *Econometrica*, 74(1):235–267, 2006. doi: 10.1111/j.1468-0262.2006.00655.x.

Alberto Abadie and Guido W. Imbens. Bias-corrected matching estimators for average treatment effects. *Journal of Business & Economic Statistics*, 29(1):1–11, 2011. doi: 10.1198/jbes.2009.07333.

Shivani Agarwal and Partha Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10(16):441–474, 2009. URL https://jmlr.org/papers/v10/agarwal09a.html.

Rachael C. Aikens, Dylan Greaves, and Michael Baiocchi. A pilot design for observational studies: Using abundant data thoughtfully. *Statistics in Medicine*, 39(30):4821–4840, 2020. doi: 10.1002/sim.8754. URL https://arxiv.org/abs/1908.09077.

Joseph Antonelli, Matthew Cefalu, Nathan Palmer, and Denis Agniel. Doubly robust matching estimators for high dimensional confounding adjustment. *Biometrics*, 74(4):1171–1179, 2018. doi: 10.1111/biom.12887. URL https://jantonelli111.github.io/files/biom12887.pdf.

Susan Athey, Julie Tibshirani, and Stefan Wager. Generalized random forests. *The Annals of Statistics*, 47(2):1148–1178, 2019. doi: 10.1214/18-AOS1709. URL https://arxiv.org/abs/1610.01271.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993. doi: 10.1109/18.256500.

Peter L. Bartlett, Dylan J. Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper/7204-spectrally-normalized-margin-bounds-for-neural-networks.pdf.

Ralph A. Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3–4):324–345, 1952. doi: 10.1093/biomet/39.3-4.324.

Christopher Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. Learning to rank using gradient descent. In *Proceedings of ICML*, pp. 89–96, 2005.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1):266–298, 2010. doi: 10.1214/09-AOAS285. URL https://www.acadiau.ca/~hchipman/papers/AOAS285.pdf.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://papers.nips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of U-statistics. *The Annals of Statistics*, 36(2):844–874, 2008. doi: 10.1214/009052607000000910. URL https://www.stat.uchicago.edu/~lekheng/meetings/mathofranking/ref/vayatis.pdf.

Stéphan Clémençon, Igor Colin, and Aurélien Bellet. Scaling-up empirical risk minimization: Optimization of incomplete U-statistics. *Journal of Machine Learning Research*, 17(76):1–36, 2016. URL https://jmlr.org/papers/v17/15-012.html.

Oscar Clivio, Fabian Falck, Brieuc Lehmann, George Deligiannidis, and Chris Holmes. Neural score matching for high-dimensional causal inference. In *Proceedings of AISTATS*, volume 151 of *PMLR*, pp. 7076–7110, 2022. URL https://proceedings.mlr.press/v151/clivio22a.html.

Rajeev H. Dehejia and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association*, 94(448): 1053–1062, 1999. doi: 10.1080/01621459.1999.10473858. URL https://users.nber.org/~rdehejia/papers/dehejia_wahba_jasa.pdf.

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

Ben B. Hansen. The prognostic analogue of the propensity score. *Biometrika*, 95(2):481–488, 2008. doi: 10.1093/biomet/asn004.

Reinhard Heckel, Nihar B. Shah, Kannan Ramchandran, and Martin J. Wainwright. Active ranking from pairwise comparisons and when parametric assumptions do not help. *The Annals of Statistics*, 47(6):3099–3126, 2019. doi: 10.1214/18-AOS1772.

Daniel E. Ho, Kosuke Imai, Gary King, and Elizabeth A. Stuart. MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, 42(8):1–28, 2011. doi: 10.18637/jss.v042.i08. URL https://www.jstatsoft.org/article/view/v042i08.

Kevin G. Jamieson and Robert D. Nowak. Active ranking using pairwise comparisons. In *Advances in Neural Information Processing Systems*, volume 24, 2011. URL https://arxiv.org/abs/1109.3701.

Thorsten Joachims. Optimizing search engines using clickthrough data. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 133–142, 2002. doi: 10.1145/775047.775067. URL https://www.cs.cornell.edu/~tj/publications/joachims_02c.pdf.

Quinn Lanners, Harsh Parikh, Alexander Volfovsky, Cynthia Rudin, and David Page. Variable importance matching for causal inference. In *Proceedings of UAI*, volume 216 of *PMLR*, pp. 1174–1184, 2023. URL https://proceedings.mlr.press/v216/lanners23a.html.

Finbarr P. Leacy and Elizabeth A. Stuart. On the joint use of propensity and prognostic scores in estimation of the average treatment effect on the treated: A simulation study. *Statistics in Medicine*, 33(20):3488–3508, 2014. doi: 10.1002/sim.6030. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3995901/.

Sheng Li and Yun Fu. Matching on balanced nonlinear representations for treatment effects estimation. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/b2eeb7362ef83deff5c7813a67e14f0a-Abstract.html.

Lokesh Nagalapatti, Pranava Singhal, Avishek Ghosh, and Sunita Sarawagi. PairNet: Training with observed pairs to estimate individual treatment effect. In *Proceedings of ICML*, volume 235 of *PMLR*, pp. 37237–37259, 2024. URL https://proceedings.mlr.press/v235/nagalapatti24a.html.

Harsh Parikh, Cynthia Rudin, and Alexander Volfovsky. MALTS: Matching after learning to stretch. *Journal of Machine Learning Research*, 23(240):1–42, 2022. URL https://jmlr.org/papers/v23/21-0053.html.

Vikas C. Raykar, Harald Steck, Balaji Krishnapuram, Cary Dehing-Oberije, and Philippe Lambin. On ranking in survival analysis: Bounds on the concordance index. In *Advances in Neural Information Processing Systems*, volume 20, 2007. URL https://papers.nips.cc/paper_files/paper/2007/file/33e8075e9970de0cfea955afd4644bb2-Paper.pdf.

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. doi: 10.1093/biomet/70.1.41.

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. URL https://proceedings.mlr.press/v70/shalit17a.html.

Elizabeth A. Stuart. Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21, 2010. doi: 10.1214/09-STS313. URL https://arxiv.org/abs/1010.5586.

Yichong Xu, Sivaraman Balakrishnan, Aarti Singh, and Artur Dubrawski. Regression with comparisons: Escaping the curse of dimensionality with ordinal information. *Journal of Machine Learning Research*, 21(162):1–54, 2020. URL https://jmlr.org/papers/v21/19-505.html.

A PROOFS AND LEARNING ASSUMPTIONS

A.1 COMPARISON CALIBRATION AND NOISE BOUNDARIES

For Proposition 1, if G is standard Gumbel, $E = e^{-G}$ is exponential with rate one. Let $a = \exp\{[m_0(x) - m_0(x')]/s\}$. The event $Y > Y'$ is equivalent to $E < aE'$, and

$$\mathbb{P}(E < aE') = \int_0^\infty (1 - e^{-av})e^{-v}\, dv = \frac{a}{1+a}.$$

Conditional Bernoulli cross-entropy is uniquely minimized at the true probability, hence at logit $d_* = [m_0(x) - m_0(x')]/s$. For any minimizing score, $e = g - g_*$ obeys $e(X) = e(X')$ almost surely. Because the two covariates are independent, this product-almost-sure equality forces e to be constant almost surely, proving identification up to a constant without an additional second-moment assumption on the target.

For iid errors independent of covariates, with finite mean zero and a common continuous law, $\epsilon' - \epsilon$ is symmetric even when the individual errors are asymmetric. A strictly increasing difference CDF implies that $\mathbb{P}(Y > Y' \mid x, x') > 1/2$ exactly when $m_0(x) > m_0(x')$. This establishes order, not a logistic link or affine score calibration. Adding arbitrarily small independent continuous jitter to the mean-order reversal in the main text removes ties without changing the reversal. Thus unrestricted heteroskedastic or skewed noise is not covered by the prognosis-calibration statement.

Noiseless labels $\mathbf{1}\{m_0(x) > m_0(x')\}$ identify only order: every increasing transformation preserves them. Separability can also prevent an unconstrained logistic objective from attaining a finite minimum. Our oracle-label network is regularized and bounded, but its fitted distances remain different from numeric oracle prognosis.

A.2 PATIENT-SCALE LEARNING AND SAMPLED COMPARISONS

Learning at patient scale. Let $\mathcal{G}$ be a fixed finite class with $|g| \le B$ for every member, containing a bounded centered version of g_*. Let U_n be the average loss over all pairs of n independent historical controls, and let R be population logistic risk. Put

$$\kappa = \mathrm{sigmoid}(2B)\{1 - \mathrm{sigmoid}(2B)\}, \quad L = \log(1 + e^{2B}), \quad a_n = \left(\frac{4}{\kappa} + \frac{4L}{3}\right)\frac{\log(|\mathcal{G}|/\delta)}{\lfloor n/2 \rfloor}.$$

An exact minimizer $\widehat{g}$ of U_n satisfies, with probability at least $1 - \delta$,

$$R(\widehat{g}) - R(g_*) \le a_n, \qquad \mathbb{E}_P[e_{\widehat{g}}(X)^2] \le a_n/\kappa, \quad e_g = g - g_* - \mathbb{E}_P(g - g_*). \tag{6}$$

Generalization for ranking has U-statistic and algorithmic-stability analyses (Clémençon et al., 2008; Agarwal & Niyogi, 2009); incomplete U-statistics already separate patient sampling from computational pair sampling (Clémençon et al., 2016). The n^{-1} finite-class bound uses a permutation reduction of the U-statistic and a Bernstein condition for the logistic excess loss, following established ranking machinery (Clémençon et al., 2008). Its sample size is the number of patients, not the number of pairs. The following argument proves the constants and gives a sampled-pair version with an explicit $M^{-1/2}$ approximation term and optimization gap. The finite-class and realizability assumptions are not a certificate for the fitted continuous neural class.

We give the complete finite-class argument. The class $\mathcal{G}$ is deterministic, every candidate and the target share bound B, the Gumbel model is correct, and a centered target belongs to the class. For a candidate g, write $Q_g = \ell_g - \ell_{g_*}$ and $\Delta_g = R(g) - R(g_*)$. Conditional logistic curvature is at least κ on the common logit interval $[-2B, 2B]$. Therefore

$$\Delta_g \ge \frac{\kappa}{2}\mathbb{E}(d_g - d_*)^2 = \kappa \mathbb{E}_P e_g^2. \tag{7}$$

The pointwise loss is 1-Lipschitz in its logit, so

$$\mathrm{Var}(Q_g) \le \mathbb{E}Q_g^2 \le \mathbb{E}(d_g - d_*)^2 \le 2\Delta_g/\kappa.$$

Also $Q_g \in [-L, L]$ and $0 \le \Delta_g \le L$. Put $m = \lfloor n/2 \rfloor$. The complete U-statistic is an average over permutations of averages of m disjoint-pair kernels. Each permuted average contains independent

kernels. Jensen's inequality for the exponential transfers the ordinary Bernstein bound to the U-statistic (Clémençon et al., 2008, Appendix A):

$$\mathbb{P}\{U_n(Q_g) - \Delta_g \le -t\} \le \exp\left\{-\frac{mt^2}{2\,\mathrm{Var}(Q_g) + (4L/3)t}\right\}. \tag{8}$$

Taking $t = \Delta_g > 0$ bounds $\mathbb{P}\{U_n(Q_g) \le 0\}$ by $\exp\{-m\Delta_g/(4/\kappa + 4L/3)\}$. An exact ERM obeys $U_n(Q_{\widehat{g}}) \le 0$. A union bound over candidates with excess larger than a_n proves (6); equation (7) gives the centered-score error. The proof counts independent patients, although all their pairs are used.

Sampled pairs and optimization. Conditional on historical data, sample M nonself pairs uniformly with replacement. The symmetric logistic kernel gives the same average for ordered and unordered sampling when outcomes have no ties. For $\delta_p \in (0, 1)$, let

$$t_M = L\sqrt{\frac{\log(2|\mathcal{G}|/\delta_p)}{2M}}.$$

Conditional Hoeffding concentration and a union bound give $\sup_g |\widetilde{U}_M(g) - U_n(g)| \le t_M$ with conditional probability at least $1 - \delta_p$. If the selected score is within $\zeta \ge 0$ of the sampled empirical minimum, then $U_n(\widehat{g}_M) - U_n(g_*) \le 2t_M + \zeta$.

Using (8) with $t = \Delta_g/2$ shows that, simultaneously for all scores whose complete empirical excess is at most η,

$$\Delta_g \le \max\left\{2\eta, \left(\frac{16}{\kappa} + \frac{8L}{3}\right)\frac{\log(|\mathcal{G}|/\delta_h)}{\lfloor n/2\rfloor}\right\}$$

with historical probability at least $1 - \delta_h$. Set $\eta = 2t_M + \zeta$ and combine the events. With probability at least $1 - \delta_h - \delta_p$,

$$\mathbb{E}_P e^2_{\widehat{g}_M} \le \frac{b_{n,M,\zeta}}{\kappa}, \quad b_{n,M,\zeta} = \max\left\{4t_M + 2\zeta, \left(\frac{16}{\kappa} + \frac{8L}{3}\right)\frac{\log(|\mathcal{G}|/\delta_h)}{\lfloor n/2\rfloor}\right\}. \tag{9}$$

This bound can replace ε^2 in Theorem 1. More sampled pairs reduce the pair-approximation term without creating new independent patients. The bound does not certify an optimization gap for early-stopped Adam, a continuous neural-class complexity, or realizability of an unbounded prognosis with a bounded network. Those are separate assumptions, not conclusions of the experiment.

Approximation and contamination. If g_* shares the bound B but is outside $\mathcal{G}$, uniform U-statistic Hoeffding concentration gives the slower bound

$$\Delta_{\widehat{g}} \le \inf_{g\in\mathcal{G}} \Delta_g + 2L\sqrt{\frac{\log(2|\mathcal{G}|/\delta)}{2\lfloor n/2\rfloor}}.$$

This includes approximation error but still assumes the correct Gumbel calibration. For outcome replacements in Proposition 2, the fraction of pairs touching at least one changed unit is $q_{n,k}$. Each label change perturbs the loss by $|d_g| \le 2B$. Comparing the changed minimizer and the original minimizer through their two objective perturbations gives $4Bq_{n,k}$. Combining the preceding empirical-excess argument with $\eta = 4Bq_{n,k}$ yields a bound relative to the clean target. It does not assert that arbitrary contaminated means equal that target.

A.3 Matching MSE and the adaptive-donor transfer

Exact MSE. Conditional on the historical fit and future (X, A), let $v_{a,i}$ be the conditional variance of $Y_i(a)$, and assume independent future units. For any outcome-independent donor map,

$$\mathbb{E}[(\widehat{\tau}_J - \tau_X)^2 \mid H, X, A] = \left[\frac{1}{N_1}\sum_{i\in\mathcal{T}}\{m_0(X_i) - m_0(X_{J(i)})\}\right]^2 + \frac{\sum_{i\in\mathcal{T}} v_{1,i}}{N_1^2} + \frac{\sum_{j\in\mathcal{C}} K_j(J)^2 v_{0,j}}{N_1^2}. \tag{10}$$

The matching discrepancy and reused-control variance are distinct. This is the usual matching decomposition (Abadie & Imbens, 2006); our use of MSE does not require a new inference procedure. Bias-corrected matching is a separate established extension (Abadie & Imbens, 2011).

Condition on historical fitting, fixed algorithmic randomness, and future (X, A). Decompose each observed future outcome into its conditional mean and a mean-zero residual. After subtracting τ_X, the deterministic term is the average signed untreated-prognosis discrepancy. Grouping donor residuals by unit gives

$$\widehat{\tau}_J - \tau_X = \frac{1}{N_1}\sum_{i\in\mathcal{T}}\{m_0(X_i) - m_0(X_{J(i)})\} + \frac{1}{N_1}\left(\sum_{i\in\mathcal{T}}\epsilon_{1,i} - \sum_{j\in\mathcal{C}}K_j\epsilon_{0,j}\right).$$

Independent units give zero cross-covariances, proving (10). Potential outcomes within the same unit need not be independent: only one arm is observed per unit. With heterogeneous effects the target remains τ_X; additional variability would be needed for a superpopulation target.

To prove Theorem 1, let $D_0(J)$ and $D_g(J)$ be mean squared pair differences in true prognosis and fitted score. Let $E_e(J)$ denote the left side of (5). For every capped map, $(a-b)^2 \le 2a^2 + 2b^2$ and $K_j \le r$ prove (5). Density-ratio bounds consequently give

$$\mathbb{E}_F E_e(J) \le 2\left(\omega_1 + \frac{rN_0}{N_1}\omega_0\right)\varepsilon^2 = 2C_r\varepsilon^2,$$

even for adaptively selected J. Let J_r^* minimize D_0 over the same capped maps. Centering score error does not change any pair difference, and $m_0 = sg_* + c$. The triangle inequality followed by score-assignment optimality yields

$$\begin{aligned}\sqrt{D_0(\widehat{J}_r)} &\le s\sqrt{D_g(\widehat{J}_r)} + s\sqrt{E_e(\widehat{J}_r)} \\ &\le \sqrt{D_0(J_r^*)} + s\sqrt{E_e(J_r^*)} + s\sqrt{E_e(\widehat{J}_r)}.\end{aligned}$$

Applying Minkowski's inequality over the future cohort gives $\sqrt{\mathbb{E}_F D_0(\widehat{J}_r)} \le \sqrt{A_r} + 2s\sqrt{2C_r\varepsilon^2}$. The squared signed mean discrepancy in (10) is bounded by D_0. Finally, $\sum_j K_j^2 \le r\sum_j K_j = rN_1$, so the variance terms are at most $v(1+r)/N_1$. Combining these facts proves (4). The oracle objective A_r concerns squared prognosis gaps; it is not the minimum achievable MSE over all assignments.

B Implementation and complete comparisons

Neural fitting. Each network has $32p + 1121$ parameters: 1441 at $p = 10$ and 1601 at $p = 15$. RankNet uses a bounded output $8\tanh(h/8)$; outcome regressors use a linear output. Thus equal hidden layers and parameter counts do not isolate the loss alone: the output constraint also differs. Hidden weights are initialized identically across neural methods in each repetition. All use Adam with learning rate 0.001, weight decay 0.001, and at most 800 gradient updates. Batches contain 128 patients for regression or 256 pairs for ranking. RankNet draws 6000 fixed nonself training pairs uniformly with replacement; validation uses 4000 pairs from separate controls. Regression outcomes are centered by their fitting median and scaled by their fitting IQR, bounded below by 0.1. Huber's transition is 1 on this scale.

Validation is evaluated every ten updates using the method's prespecified objective. The earliest best checkpoint is retained; stopping requires at least 100 updates and no improvement for 80 updates. Neither future outcomes nor true prognosis select feasible learners. Median selected updates are 215 for squared-loss MLP, 290 for Huber MLP, and 250 for RankNet. The common maximum update count does not make patient and pair gradients statistically equivalent; the comparison controls architecture and selection access rather than claiming identical optimization difficulty.

Other learners (six-setting synthetic experiment). Histogram boosting uses 150 iterations, learning rate 0.05, at most 15 leaves, minimum 20 observations per leaf, regularization 1, and no internal early stopping. MALTS uses the supplied original implementation's hard-neighbor within-arm objective with $C = 1$ and $k = 10$, unit initial stretches, and COBYLA tolerance 0.001 and maximum

Table 3: RankNet minus comparator conditional design MSE under cap 5 (distinct from the realized-error endpoint in Figure 2). Negative values favor RankNet. Parentheses give paired Monte Carlo SE; these are simulation precision measures, not treatment-effect confidence intervals.

Family	History	MLP squared	MLP Huber	Boost absolute	MALTS
Linear	Clean	−.0003 (.0016)	.0008 (.0008)	−.0340 (.0087)	−.0275 (.0045)
Linear	Contaminated	−.1764 (.0612)	.0113 (.0055)	−.0102 (.0088)	−.0480 (.0157)
Folded	Clean	.6852 (.2190)	.9325 (.2085)	−.2214 (.2196)	−.4424 (.1740)
Folded	Contaminated	−.8633 (.3652)	−.0822 (.3683)	1.5335 (.3532)	.8057 (.3641)
Nonlinear	Clean	−.0097 (.0199)	−.0035 (.0171)	−.0119 (.0114)	.0076 (.0189)
Nonlinear	Contaminated	−.0258 (.0234)	−.0140 (.0233)	.0060 (.0194)	.0266 (.0252)

5000 evaluations. All 120 optimizer statuses report success, with at most 222 evaluations. This is the implemented original objective, not an assertion that its nonsmooth objective reached a global minimum. All covariates are continuous. Learned stretches are applied in raw covariate coordinates, consistent with the source implementation.

Data independence and uncertainty (six-setting synthetic experiment). Historical fitting, historical validation and future units are independent. Clean and contaminated runs share covariates, treatment assignments, and future noise within each family/seed. Treatment effects are constant at 1. Conditional design MSE is the squared average true prognosis mismatch plus $0.25/N_1 + 0.25\sum_j K_j^2/N_1^2$. Standard errors are the sample standard deviation across twenty repetitions divided by $\sqrt{20}$; paired comparisons use the corresponding per-repetition differences.

B.1 Sample-size fitting details

Sample-size design. We fixed $X \sim N(0, I_{10})$, $m_0(x) = 4|u_4| + 0.5x_{10}$, and propensity logit $2u_4 + 2(|u_4| - 0.8)$, clipped after the logistic transform as before. Five covariates enter prognosis and four enter propensity; prognosis depends on two indices. Treatment concentrates on the positive projection branch, while controls remain available on the negative branch. Opposite projection values share the folded component of prognosis, so a scalar learner can connect covariate-distant patients; a diagonal stretch cannot generally express this identification. Twenty confirmation seeds, disjoint from development, use total historical fitting sizes $n \in \{300, 600, 1200, 2400\}$, validation size $n/2$, and the same independent 800-patient future cohort across sizes within each seed. Nested fitting samples isolate additional historical information at fixed future donor availability. Validation patients and the sampled-pair budget also increase with n.

RankNet uses width 32 (1441 parameters), with widths 8 and 64 (169 and 4929 parameters) as ablations. Neural fits use up to 4000 updates and independent validation stopping. Ranking outputs are bounded; regression outputs are linear. MALTS fits 10 stretches using three predetermined starts, selected by historical objective. Five additional 20-seed perturbations at $n = 2400$ change one factor: projection onto 1 or 8 covariates, historical Gumbel noise scale 1.0, or propensity-logit multiplier 0.5 or 1.5. The resulting true prognosis/propensity active counts are 2/1 or 9/8 for the projection changes, and 5/4 otherwise. The prognosis remains a two-index function. All 180 designs and both matching rules are retained.

Expanded fitting protocol. The sample-size study replaces the original 800-update budget by a common 4000-update maximum. Validation is evaluated every 25 updates; stopping requires at least 500 updates and 400 updates without improvement. Training uses $M = \max(6000, 20n_C)$ sampled pairs and validation $\max(4000, 10n_{C,V})$, where $n_C, n_{C,V}$ count fitting and validation controls. Architecture, outcome scaling, Adam settings, batch sizes, bounded ranking output and linear regression output remain as specified above. Actual fitting-control counts, pair indices, validation histories, selected/stopped updates and optimizer diagnostics are saved. The change in optimization budget prevents direct pooling with the original experiment.

MALTS uses three fixed starts: unit stretches and two vectors drawn coordinatewise as $\exp\{N(0, 0.5^2)\}$. Each uses the original objective and at most 5000 evaluations. The smallest finite historical objective selects the fit, with earliest-start tie breaking. Multiple starts do not certify a global minimum.

B.2 FUTURE-OUTCOME ATT REFERENCES

The reference estimators use the same realized future outcomes and covariates as the matching estimates. Historical matching geometry is frozen before these outcomes are used. We use `dbarts` 0.9-32 for BART (Chipman et al., 2010), with 200 trees, 1000 burn-in iterations and 1000 retained posterior draws, one chain and one thread. The model is fitted only on future controls. If $\widehat{m}_{0,B}$ is its posterior mean prediction, the point estimate is $N_1^{-1}\sum_{i\in\mathcal{T}}\{Y_i - \widehat{m}_{0,B}(X_i)\}$. Saved posterior draws reproduce the reported predictions; this is not a claim of posterior convergence or interval validity.

Causal forest uses `grf` 2.5.0 (Athey et al., 2019), with 2000 trees, honest splitting, one thread and out-of-bag nuisance predictions. We call `average_treatment_effect` with treated target and AIPW. Writing its outcome, propensity and treatment-effect predictions as $\widehat{q}_i, \widehat{e}_i, \widehat{t}_i$, set $\widehat{m}_{0,i} = \widehat{q}_i - \widehat{e}_i\widehat{t}_i$ and $w_i = \widehat{e}_i/(1-\widehat{e}_i)$. The returned point estimate is the treated mean of $Y_i - \widehat{m}_{0,i}$ minus the control weighted mean of $Y_i - \widehat{m}_{0,i}$, normalized by $\sum_{i\in\mathcal{C}} w_i$. No propensity clipping or replicate removal is applied. This population-ATT procedure is evaluated against the simulation's sample conditional target; finite-sample conditional unbiasedness is not assumed.

The six synthetic settings share 60 distinct future cohorts (three families, twenty seeds); clean and contaminated historical conditions reuse each cohort. Thus future-fitted reference values repeat across historical noise conditions. Matching-only conditional design MSE integrates future errors analytically, but cannot be substituted for realized squared errors of these outcome-adaptive references. All primary reference comparisons therefore use realized squared error and twenty-repetition MCSE.

B.3 LALONDE GENERATOR, SCALE STUDY AND COMPLETE COMPARISON

The data supply fixed covariates and treatment labels: age, education, Black and Hispanic indicators, married, no degree, and 1974/1975 earnings in thousands of dollars. A histogram gradient booster fitted to the original 429 control earnings in 1978 (in thousands) defines $m_0(x) = 1000\max\{1, \widehat{f}(x)\}$. It uses 200 iterations, learning rate 0.05, at most 15 leaves, minimum 20 observations per leaf, no regularization or early stopping, and seed 9000. This generator is fixed once using the entire original control pool as a simulation-design step; experimental learners receive freshly generated outcomes, not the real earnings used to construct it.

Set $\tau(x) = 1000\{1 + 0.2z_{\text{educ}}(x)\}$, using the empirical mean and population-standard-deviation convention over all 614 education values. The potential outcomes are $Y(0) = m_0(X) + \epsilon$ and $Y(1) = Y(0) + \tau(X)$, with independent $\epsilon \sim \text{Uniform}[-1000, 1000]$. The effects are positive, so both outcomes are nonnegative. The generator does not reproduce the original earnings point mass at zero. Uniform noise also departs from the Gumbel calibration assumption; this is an empirical scale-sensitivity test, not a verification of A5.

For seeds 9000–9019, separately permute each treatment arm and split without replacement into fitting (74 treated/172 controls), validation (18/43), and future (93/214). Fitting, validation and future patient identifiers are disjoint within each repetition. Repetitions vary splits and generated errors conditional on the same finite pool; the cohorts are not independent superpopulation samples. The target is the mean $\tau(X)$ among the 93 future treated patients, recalculated in every repetition.

For scale-sensitive learners, transform historical fitting and validation outcomes to either Y in dollars or $\log(1+Y)$, keeping future outcomes in dollars. Neural learners use the expanded 4000-update protocol in Appendix B.1. The five numerical regression learners each fit both scales; MALTS fits both scales using standardized covariates and historical all-arm median/IQR outcome normalization (IQR floor 0.1), three fixed starts, and the supplied within-arm objective. Binary indicators are encoded numerically in the continuous distance. The literal coordinate-difference-then-weight arithmetic preserves the supplied objective's strict-threshold treatment of distance ties. The three-start selection uses only historical objectives.

All pairwise labels among fitting and validation controls agree exactly between Y and $\log(1+Y)$. With unchanged pair draws, initialization and rank validation, a single RankNet fit represents both scales; no separate refitting noise is introduced. Oracle-label RankNet is likewise shared. Mahalanobis, propensity matching and numeric oracle prognosis do not use the transformed historical

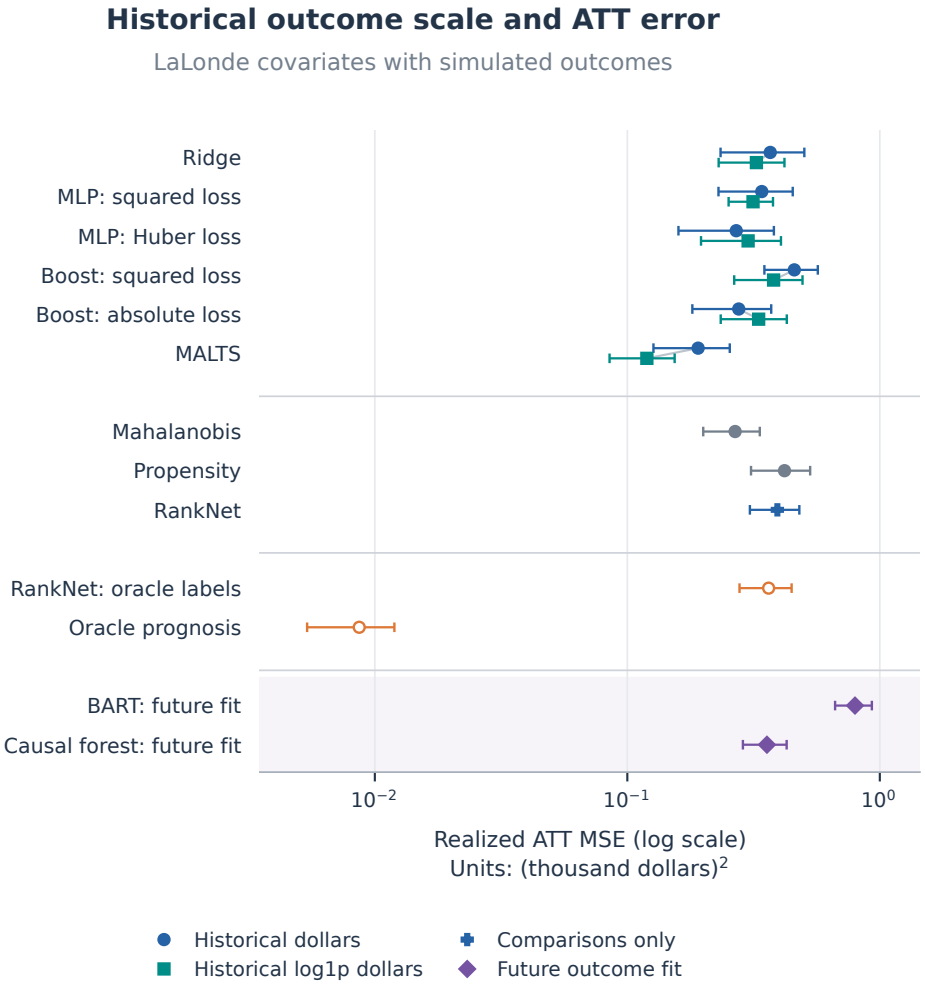

Figure 4: Historical-scale invariance does not guarantee the lowest ATT error. Semi-synthetic LaLonde outcomes have education-dependent effects; all methods are evaluated on original future dollar outcomes. Points show mean realized squared ATT error in squared thousands of dollars and one MCSE over twenty splits. Historical matching uses cap 5. Connected dollar/log points change only historical outcome supervision; RankNet is shared because its comparison labels are identical. BART and causal forest use future outcomes. Oracle methods are unavailable.

outcome values. This gives seventeen matching variants, each evaluated under cap 5 and 1NN, plus two future-outcome references: 720 estimates over twenty repetitions. The primary figure displays cap 5 and both reference estimators. Log-outcome regression learns a different target from dollar-outcome regression; its score is used directly for matching, without claiming an unbiased dollar retransformation.